

ESTIMATING DENSITY FUNCTIONS: A CONSTRAINED MAXIMUM LIKELIHOOD APPROACH [†]

Michael X. Dong and Roger J-B Wets

Department of Mathematics
University of California, Davis

in *Nonparametric Statistics*, 12(2000),549-595.

Abstract. We propose estimating density functions by means of a constrained optimization problem whose criterion function is the maximum likelihood function, and whose constraints model any (prior) information that might be available. The asymptotic justification for such an approach relies on the theory of epi-convergence. A simple numerical example is used to signal the potential of such an approach.

Key Words: constrained maximum likelihood estimation, consistency, epi-convergence, Mosco-epi-convergence, ρ -epi-distance.

[†] Research supported in part by a grant of the National Science Foundation

The problem is to find an estimate \hat{h} of a density $h^0 : \Xi \rightarrow \mathbb{R}_+$ associated with a random variable ξ , given iid observations $\xi^1, \xi^2, \dots, \xi^\nu$, and any prior information that might be available about the random phenomena modeled by ξ .

Quite a number of procedures have been suggested to deal with this primary statistical question. They come in two basic flavors: *parametric* and *nonparametric estimation*. In the parametric case, the prior information allows us to single out a specific class of density functions characterized by a parameter $\theta \in \mathbb{R}^N$. Estimating h^0 is then reduced to finding a best estimate for this parameter. In some sense, it's a density estimation problem with nearly complete information. In the nonparametric case, no prior information is available except that the distribution of the random phenomenon can be described in terms of a density function. Now, one has to find a function h whose only known property is that it is a density function. These two problem types are in some sense at the opposite ends of the class of problems of that actually fit under the "density estimation" label. In practice, some partial information is usually available about the unknown density, but not quite enough to be able to pinpoint the parametric class to which h^0 belongs. For example, one might know (or suspect) that h^0 is unimodal, or one might have (or stipulate) bounds on certain quantiles, or still one might even know (or suspect) that h^0 belongs to a neighborhood of a certain density \tilde{h} . This partial information about h^0 should play an important role in restricting the choice of \hat{h} to certain subclasses of density functions, especially when the sample size is relatively small, i.e., doesn't quite reach the asymptotic range.

The standard approach to nonparametric estimation, say kernel estimation [31, 32], has some shortcomings that are difficult to patch, at least at the theoretical and computational levels. In particular, it isn't really possible to include prior information, e.g., bounds on moments or the support, level of smoothness, unimodality, and so on. Moreover, it's also difficult to say in which sense the resulting estimated density is the "best" possible given the information available; in fact, it usually isn't. Finally, when only a small number of samples have been collected, relying on these standard estimation procedures could be disastrous.

Another approach, that is going to be followed in this paper, is to formulate the density estimation problem as an optimization problem: find a function \hat{h} that maximizes the maximum likelihood function; note however that the techniques developed here aren't limited to this particular loss function. The choice of \hat{h} is always subject to the constraints $\hat{h} \geq 0$ and $\int \hat{h}(t) dt = 1$ that identify \hat{h} as a density function. If these are the only constraints included in the formulation of the optimization problem, the answer turns out to be somewhat meaningless, at least in general. The optimal \hat{h} is then the summation of Dirac functions that assigns equal mass to each sample point; the counterpart of the empirical measure. Usually, however more information is available about the distribution of the underlying stochastic phenomena. For example, it may be known that h^0 is a smooth function, thus it would be natural to restrict the choice of an estimator to functions with prescribed smoothness properties, i.e., introduce additional constraints in the optimization

problem. Similarly, any piece of information that might be available about h^0 will be reflected by additional constraints to be imposed on the choice of \hat{h} : unimodality, bounds on some moments, and so on.

“Prior” information, *including* modeling assumptions, is extremely valuable when the samples at hand are too few to reach the asymptotic range, *as is almost always the case* when $\boldsymbol{\xi}$ is \mathbb{R}^d -valued with $d \geq 2$. Indeed, the prior information constitutes a larger share of the total information available when only a small number of samples have been collected. In [38], Samaniego and Reneau express similar concerns about including prior information and modeling assumptions in the formulation of the estimation problem. In terms of our optimization problem this means that the constraints that describe the prior information will then, as they should, play a more significant role in determining the optimal estimate.

There have been some attempts in the literature to include smoothness considerations in the calculation of the estimator in the form of a *single* constraint, and this constraint has been included in the formulation of the estimation problem in the form of a penalty term with a coefficient to be adjusted “in practice”. The relationship between our model and this literature is clarified in section §4.

The quality of a statistical estimator can’t be determined exclusively by its asymptotic properties. Implicit in our approach is the postulate that only for estimators that are “best” in some sense, and that use all the information available (which includes a finite sample), is it possible to find a convincing practical justification. It will be shown in this paper that also a theoretical justification is at hand. Although, Thompson and Tapia don’t quite deal with as general a formulation of the estimation problem as that considered in this paper, their beautiful monograph [44] on “Nonparametric Function Estimation, Modeling and Simulation,” in some way a forerunner of this work, shares with it the same underlying strategy. Other work, of possibly more direct antecedence is that dealing with the consistency of the solutions of stochastic optimization problems [14, 24, 40, 29, 35] and certain constrained estimation problems [36, 49, 45, 50, 13, 17, 41].

The density estimation problem is formulated, and the argmin estimator is introduced in §1 which also provides a number of illustrative examples. Section 2 outlines our approach which basically consists in viewing the estimation problems as a sequence optimization problems that converge, in a sense to be made precise, to a limit problem whose optimal solution is the true density as explained in §3. The basic tools that are going to be employed are laid out in §5 and §6, with the consistency results coming in §7. The approach is illustrated in §8 by a simple numerical example. An appendix reworks the proof in [1] of a law of large numbers for random lsc functions in order to relax slightly a couple of assumptions. Convergence rates for the argmin-estimators can be calculated. This will be dealt with in a separate paper.

1. The argmin estimator

Let $\xi^1, \xi^2, \dots, \xi^\nu$ be independent observations of a random variable $\boldsymbol{\xi} : (\Xi, \mathcal{A}) \rightarrow (\mathbb{R}^d, \mathcal{B})$, $\Xi \subset \mathbb{R}^d$, and P^ν the empirical measure generated by these samples. This paper is concerned with the following density estimator:

$$\hat{h}^\nu \in \operatorname{argmin} \{E^\nu L_0(h) \mid h \in S \subset H\}, \quad (1.1)$$

where H is some functional space, L_0 is a criterion function, $E^\nu L_0$ is an expectation functional,

$$E^\nu L_0(h) := \int_{\Xi} L_0(\xi, h) P^\nu(d\xi),$$

and

$$S = \{h \in A \subset H \mid \int_{\Xi} h(\xi) d\xi = 1, h(\xi) \geq 0, \forall \xi \in \Xi\} \quad (1.2)$$

where the set A consists of those functions h that satisfy whatever additional information (prior information or modeling assumptions) that might be available about the density function that is to be estimated. It will be assumed that such a \hat{h}^ν exists; although existence and uniqueness aren't going to be of major concern here, they will be obtained for some important special cases.

When L_0 is the maximum likelihood function,

$$L_0(\xi, h) = \begin{cases} -\ln h(\xi) & \text{if } h(\xi) > 0, \\ \infty & \text{otherwise,} \end{cases}$$

one refers to the optimization problem (1.1) as the *constrained maximum likelihood density estimation problem*. Since most of the criterion functions that one might want to use in the formulation of the density estimation problem lead naturally to minimization, and since most the optimization literature is usually presented in terms of a canonical minimization problem, it will be convenient to also formulate the ‘‘maximum’’ likelihood estimation as a minimization problem.

It will be shown that \hat{h}^ν is a consistent estimator; in [12], we obtain, via the theory of large deviations, some results about the convergence rate of \hat{h}^ν to the ‘‘true’’ density. Of course, the theory doesn't depend on having prior information, so the ‘‘usual’’ framework is included as a special case. However, such additional information that often has been ignored, because statistical theory didn't validate its use, can now be part of the formulation of the estimation problem. A few simple, illustrative examples follow:

Example 1.1. Smoothness. Let $\boldsymbol{\xi}$ be real-valued. Suppose there is some justification for insisting on ‘‘smoothness’’ of the estimator. Assuming the search for an estimator is restricted to differentiable functions, one could impose the following constraint:

$$\int \frac{h'(\xi)^2}{h(\xi)} d\xi \leq \beta,$$

where the term on the left is the *Fisher information*. The connection between the resulting estimator and the penalized maximum likelihood estimator will be clarified in §4.

If such a constraint is included in the formulation of the estimation problem, it implicitly means that this bound β is known, or at least, that such a constraint can be accepted as a reasonable modeling assumption. Of course, nothing would prevent a post-solution analysis of the dependence of the estimates on β . The theory of *sieves estimates* [16, 42] is concerned with the discovery of the appropriate level of smoothness to include in the formulation of the the estimation problem. Although we shall deal with this issue to some extent in [11], here we view a ‘smoothness condition’ as just one of the many type of constraints that might, or might not, be included in the formulation of the estimation problem.

Example 1.2. *Bounds on moments.* Suppose $\boldsymbol{\xi}$ is real-valued and there is some (prior) information about the expectation and the variance of $\boldsymbol{\xi}$, namely,

$$0 \leq \mu_1 \leq E\{\boldsymbol{\xi}\} \leq \mu_2, \quad \sigma_1^2 \leq \text{var}\{\boldsymbol{\xi}\} \leq \sigma_2^2,$$

assuming that $\sigma_2^2 \geq \sigma_1^2 + (\mu_2^2 - \mu_1^2)$. Then, the set A describing the additional information consists of the functions $h \in H$ satisfying

$$\mu_1 \leq \int \xi h(\xi) d\xi \leq \mu_2, \quad \sigma_1^2 + \mu_2^2 \leq \int \xi^2 h(\xi) d\xi \leq \sigma_2^2 + \mu_1^2.$$

One can also incorporate constraints on the support as in [39].

Example 1.3. *Shape.* Usually to include shape information an infinite number of constraints will be required. For example, if it is known that the true density $h^0 : \mathbb{R}^d \rightarrow \mathbb{R}_+$ is continuous and long-concave (strongly unimodal), i.e., that the set A describing the additional information consists of the functions of type,

$$h(\xi) = e^{-Q(\xi)} \quad \text{with } Q \text{ convex,}$$

then, assuming that Q is C^2 , the constraints take the form

$$\langle z, \nabla^2 Q(\xi) z \rangle \geq 0, \quad \forall z \in \mathbb{R}^d, \forall \xi \in \mathbb{R}^d$$

where $\nabla^2 Q(\xi)$ is the Hessian of Q at ξ . This might require the use of optimization techniques developed for positive definite programming [7].

As another simple example of this ilk, consider the case when $h^0 : \mathbb{R} \rightarrow \mathbb{R}_+$ is known (or suspected) to be smooth and monotone decreasing. Then, the constraint

$$h'(\xi) \leq 0, \quad \forall \xi \in \Xi.$$

will be among those determining the set A ; cf. [19, 20]. In §8, we compare the estimates obtained when such a constraint is or is not included in the formulation of the estimation problem.

Example 1.4. *The ‘bayesian’ premise.* There are many ways to include information of bayesian type. One can, for example, restrict the estimators to the neighborhood of a certain density, say \tilde{h} . The set A then consists of the functions satisfying

$$\|h - \tilde{h}\|_A \leq \alpha, \quad \alpha > 0$$

where $\|\cdot\|_A$ is an appropriate norm. Or, one could include in the formulation of the optimization problem a penalty term that would take into account a certain probability distribution on the space of density functions centered at \tilde{h} . This will be explored in a projected paper [11].

2. Epi-consistency

The consistency of the estimators will be obtained as a consequence the consistency of the *empirical* optimization problems (1.1) (defining \hat{h}^ν) with a certain limit problem whose optimal solution (argmin) is h^0 , the true density. Since the empirical measure P^ν can be viewed as approximating the probability distribution P of $\boldsymbol{\xi}$, one might surmise that this limit problem is:

$$\begin{aligned} \min \quad EL_0(h) &= \int_{\Xi} L_0(\xi, h) P^0(d\xi), \\ h &\in S \subset H, \end{aligned} \tag{2.1}$$

where P^0 is the actual distribution of the random variable $\boldsymbol{\xi}$. Assuming that h^0 actually solves this latter optimization problem, the consistency of the estimated densities \hat{h}^ν would be established if the consistency of the empirical optimization problems implied the consistency of their optimal solutions. This means that the notion of consistency must be based on a convergence notion that implies the convergence of the optimal solutions of optimization problems.

The theory of epi-convergence had its origin in the development of an approximation theory for variational problems; cf. [52, 2, 34, 4, 6] and is mostly concerned with being able to assert that the solution of an approximating problem provides a reasonable approximation to the solution of a certain limit problem. The framework in which the results are usually formulated is as follows: Every optimization problem, say

$$\min f_0(x) \text{ such that } f_i(x) \leq 0, i \in I_1, f_i(x) = 0, i \in I_2, \quad x \in S \subset X,$$

can equivalently be expressed as the minimization of just one function, provided one allows this function to take on values in the extended reals (or at least, $(-\infty, \infty]$). Indeed, with

$$f(x) = \begin{cases} f_0(x) & \text{if } f_i(x) \leq 0, i \in I_1, f_i(x) = 0, i \in I_2, x \in S; \\ \infty & \text{otherwise;} \end{cases}$$

the optimization problem

$$\min f(x), \quad x \in X$$

has the same (optimal) solutions and optimal value as the original problem. The function f is sometimes called the *essential objective*. At the conceptual and theoretical level this device is very powerful since it allows us to identify each optimization problem with just one function. In particular, the question of approximating optimization problems becomes then one of approximating (extended real-valued) functions. Moreover, when S is closed and the functions f_0, f_i are continuous (or just lower semicontinuous), the function f is lower semicontinuous, and consequently the attention can be focused on lower semicontinuous (lsc) functions, although the theory of epi-convergence itself doesn't need such a restriction.

The primary result of the theory of epi-convergence now takes the following form: If the f^ν are extended real-valued functions that epi-converge to f , then the argmin f^ν converge, in a sense to be specified later, to argmin f .

This is exactly what is needed to deal with the consistency of the estimated densities. To conform to this set-up, one identifies the empirical optimization problems (1.1) with the functions $E^\nu L : H \rightarrow \overline{\mathbb{R}}$ defined by

$$E^\nu L(h) = \int_{\Xi} L(\xi, h) P^\nu(d\xi), \quad (2.2)$$

and the limit optimization problem (2.1) with the function $EL : H \rightarrow \overline{\mathbb{R}}$ defined by

$$EL(h) = \int_{\Xi} L(\xi, h) P^0(d\xi), \quad (2.3)$$

where

$$L(\xi, h) = \begin{cases} L_0(\xi, h) & \text{if } h \in S; \\ \infty & \text{otherwise.} \end{cases} \quad (2.4)$$

Roughly speaking, given the empirical measures P^ν generated from the samples ξ^1, \dots, ξ^ν , the epi-convergence of the functions $E^\nu L$ to EL would then guarantee the convergence of the estimators \hat{h}^ν to h^0 .

When dealing with consistency, however, one must take into account every possible sequence of samples, i.e., the $E^\nu L$ are actually random functions since they depend on the empirical measure \mathbf{P}^ν that in turn depends on the observations of ν iid (independent identically distributed) random variables ξ^1, \dots, ξ^ν . These random functions $E^\nu L$ are said to be (*strongly*) *epi-consistent* with EL if they epi-converge almost surely to EL . This is precisely the result that will be derived when L_0 is the maximum likelihood function, and from this will follow the (strong) consistency of the constrained maximum likelihood estimators.

Let's point out once more that this approach isn't limited to the choice of the maximum likelihood function as the criterion function. The analysis that is going to follow can

also be carried out for any other criterion function one might choose, e.g., least squares. Of course, some of the details, proof technique, and assumptions might turn out to be somewhat different, but one would end up with similar results. That's why so far, in the description of the overall approach there has only been parenthetical reference to the maximum likelihood case.

3. The limit problem

When the criterion function is the maximum likelihood function, the limit problem takes on the form:

$$EL(h) = \begin{cases} -\int_{\Xi} \ln h(\xi) P^0(d\xi) & \text{if } h \in S \subset H, \\ \infty & \text{otherwise,} \end{cases}$$

and so far, it has been taken for granted that

$$h^0 = \operatorname{argmin} \{ EL(h) \mid h \in H \}.$$

Let's now examine this in further detail. To this end, let's introduce the *Kullback-Leibler discrepancy* between two density functions h, g :

$$K(h, g) := -\int_{\Xi} \ln \left(\frac{h(\xi)}{g(\xi)} \right) g(\xi) d\xi.$$

It has the following important properties,

- (i) for all density functions h and g , $K(h, g) \geq 0$,
- (ii) for any density h , $K(h, h) = 0$,

that follow from

$$g \in \operatorname{argmax}_h \left\{ \int g(\xi) \ln h(\xi) d\xi \mid \int h(\xi) d\xi = 1, h \geq 0 \right\}.$$

With h^0 be the (true) density function associated with P^0 , the function EL defining the limit problem can also be written as,

$$EL(h) = \begin{cases} -\int_{\Xi} (\ln h(\xi)) h^0(\xi) (d\xi) & \text{if } h \in S \subset H, \\ \infty & \text{otherwise.} \end{cases}$$

After adding the constant $\int (\ln h^0(\xi)) h^0(\xi) (d\xi)$, the limit problem simply becomes

$$\min K(h, h^0) \quad \text{such that } h \in S \subset H,$$

which means that h^0 is a solution of this problem, or equivalently a solution of the limit problem ($\min EL$), as long as h^0 belongs to S . And assuming that S is closed, if $h^0 \notin S$, the limit problem picks out a density in S that is as close as possible to h^0 in terms of the Kullback-Leibler discrepancy. If the constraints that define the set S are limited to "hard" information one might have about h^0 , then h^0 will always be included in S . However, if some of these constraints come from modeling assumptions, then there is the possibility that h^0 might have been eliminated from S the set of feasible solutions of the estimation problem.

4. Relation to the literature

It's beyond the scope of this article to review of the existing literature on the maximum likelihood estimator. Consistency of the maximum likelihood estimators is well known, one can consult [37] and [48] for approximate maximum likelihood estimators. The attention here will be focussed on those articles that consider nonparametric estimation problems with some side condition(s). The variants of the maximum likelihood estimator suggested in the literature rely on adding a penalty term to the criterion function. Usually, the aim is to guarantee a certain level of smoothness for the optimal estimator.

Good and Gaskins [18] were the first to suggest estimators based on a penalized maximum likelihood criterion. Existence and uniqueness of the Good and Gaskins estimators were obtained by de Montricher, Tapia and Thompson [9], and consistency by V.K. Klonias [26]. de Montricher, Tapia and Thompson [9], and Silverman [43] proposed different penalty terms, and proved the consistency of the resulting estimators. Leonard [28] studied the maximum likelihood method from a bayesian point of view and justified the use of the penalized maximum likelihood density estimator on the basis of having access to additional information.

The lemma below will help clarify the relationship between those proposals and the argmin estimator (1.1). Let consider the two following optimization problems:

$$\begin{aligned} \min \quad & f(x), \\ & g(x) \leq \beta, \quad x \in C \subset H, \end{aligned} \tag{CP}$$

and

$$\begin{aligned} \min \quad & f(x) + \alpha g(x), \\ & x \in C \subset H, \end{aligned} \tag{UP}$$

where $f : C \rightarrow \mathbb{R}$, $g : H \rightarrow \mathbb{R}$, C is a closed subset of H , a Hilbert space, and $\alpha, \beta \in \mathbb{R}$.

The focus in these two problems is on the (single) constraint $g(x) \leq \beta$ in (CP) and the penalty term $\alpha g(x)$ in (UP). Let's refer to (CP) as a constrained optimization problem, and (UP) as an unconstrained problem, notwithstanding the fact that the constraints $x \in C$ are still part of the formulation of (UP).

It's not true, that in general, the solutions of (CP) and (UP) are identical, or even comparable, but one has the following implications:

Lemma 4.1. *Assuming the constrained optimization problem (CP) is feasible, then the following conditions are sufficient for x^c to be an optimal solution of (CP):*

$$\exists \alpha \geq 0 \quad \text{such that} \quad \begin{cases} \alpha(\beta - g(x^c)) = 0 \\ x^c \text{ solves (UP)}. \end{cases} \tag{O.C.}$$

These conditions become necessary when, in particular, C is convex, f and g are convex, g is continuous, and there is $\tilde{x} \in C$ such that $g(\tilde{x}) < \beta$.

On the other hand, assuming now that (UP) is feasible and $\alpha \geq 0$, if x^u is an optimal solution of (UP) there always exists β such that x^u is also an optimal solution of (CP) .

Proof. These properties follow immediately from well-known optimality conditions for problems of this type. It's also easy, and instructive, to write down an explicit proof. So, let's suppose that there exist α and x^c that satisfy $(O.C.)$. If x^c is not an optimal solution of (CP) there would exist $x^0 \in C$ such that

$$g(x^0) \leq \beta, \quad f(x^0) < f(x^c).$$

On the other hand, since x^c solves (UP) , one also has

$$f(x^0) + \alpha g(x^0) \geq f(x^c) + \alpha g(x^c),$$

and consequently,

$$\alpha\beta - \alpha g(x^0) < \alpha\beta - \alpha g(x^c) = 0.$$

For this to occur, one must have $\alpha > 0$ and $\beta - g(x^0) < 0$, and then, this last inequality would contradict the feasibility of x^0 .

To prove that the conditions in $(O.C.)$ are also necessary when (CP) is a convex optimization problem, and there is a $\tilde{x} \in C$ such that $g(\tilde{x}) < \beta$, let x^c be an optimal solution of (CP) , and consider the function $\theta : \mathbb{R} \rightarrow \overline{\mathbb{R}}$ defined by

$$\theta(u) = \inf_{x \in C} \{ f(x) \mid g(x) \leq \beta + u \}.$$

This is a convex function, real-valued on neighborhood of 0 as follows from the assumptions. Moreover, θ is a decreasing function of u , hence there always exists $\alpha \geq 0$ such that

$$\theta(u) \geq -\alpha u + \theta(0), \quad \forall u \in \mathbb{R}.$$

This implies that for all $u \in \mathbb{R}$,

$$\inf_{x \in C} \{ f(x) + \alpha u \mid u \geq g(x) - \beta \} \geq f(x^c).$$

Since the preceding inequality holds for all u , one must actually have

$$\forall x \in C : \quad f(x) + \alpha(g(x) - \beta) \geq f(x^c).$$

In particular, this yields

- $\alpha(g(x^c) - \beta) = 0$,
- x^c is a solution of (UP) .

Finally, if for given $\alpha \geq 0$, x^u is a solution of (UP) , then with $\beta = g(x^u)$, x^u is a feasible solution of (CP) and the pair (x^u, α) satisfies $(O.C.)$ which means that x^u is an optimal solution of (CP) . \square

Let now return the density estimation problem:

$$\begin{aligned} \min \quad & -E^\nu \{\ln h(\boldsymbol{\xi})\} = - \int_{\Xi} \ln h(\xi) P^\nu(d\xi), \\ & \int_{\Xi} h(\xi) d\xi = 1, \quad h \geq 0, \\ & \Phi(h) \leq \beta, \\ & h \in H, \end{aligned} \tag{CE}$$

where P^ν is the empirical measure and $\Phi(h) \leq \beta$ is a (single) constraint, in addition to those requiring that the solution of this estimation problem must be a density function. Let's refer to this problem as a “constrained” maximum likelihood estimation problem. As already mentioned earlier, the constraint $\Phi(h) \leq \beta$ might have been included in the formulation of the problem to guarantee a certain level of smoothness, but that's not the only possibility.

Let's also consider the associated problem:

$$\begin{aligned} \min \quad & -E^\nu \{\ln h(\boldsymbol{\xi})\} + \alpha \Phi(h) = - \int_{\Xi} \ln h(\xi) P^\nu(d\xi) + \alpha \Phi(h), \\ & \int_{\Xi} h(\xi) d\xi = 1, \quad h \geq 0, \\ & h \in H, \end{aligned} \tag{PE}$$

where the constraint $\Phi(h) \leq \beta$ has been replaced by the penalty term $\alpha \Phi(h)$ for some $\alpha \geq 0$. This problem requires finding a density function that minimizes a “penalized” maximum likelihood function; the constraints restricting the choice of $h \in H$ to a function that is a density are still part of the formulation of the problem.

And as an immediate corollary of the preceding lemma, one has:

Corollary 4.2. *Assuming the constrained estimation problem (CE) is feasible, then sufficient conditions for \hat{h}^ν to be an optimal solution of (CE) are:*

$$\exists \alpha \geq 0 \quad \text{such that} \quad \begin{cases} \alpha(\beta - \Phi(\hat{h}^\nu)) = 0 \\ \hat{h}^\nu \text{ solves } (PE) \end{cases} \tag{O.E.}$$

These conditions also become necessary when, in particular, Φ is continuous and convex, and there is a density function $\tilde{h} \in H$ such that $\Phi(\tilde{h}) < \beta$.

On the other hand, assuming now that (PE) is feasible and that $\alpha \geq 0$, if \hat{h}^ν is an optimal solution of (PE) there always exists β such that \hat{h}^ν is also an optimal solution of (CE) .

Proof. It really suffices to observe that $h \mapsto -E^\nu \{\ln h(\boldsymbol{\xi})\}: H \rightarrow \overline{\mathbb{R}}$ is convex on the convex set $D = \{h \in H \mid \int_{\Xi} h(\xi) d\xi = 1, h \geq 0\}$, and then apply the lemma. \square

The corollary highlights the close relationship between the inclusion of a constraint in the formulation of the density estimation problem or the adding of a penalty term to the objective. In all the examples in the literature, it's the objective of the maximum likelihood estimation problem that has been modified by a penalty term to enforce smoothness, i.e., the formulation of the problem is of type (PE) . When implementing such an approach, the coefficient α is adjusted so as to yield a density exhibiting certain “appropriate” properties.

From the observations made so far, it might appear that there is nothing that favors one or the other formulation, i.e., one involving the constraint $\Phi(h) \leq \beta$ or one involving the penalty term $\alpha\Phi(h)$. One might point to corollary 4.2 as supporting evidence for such an assertion. However, corollary 4.2 doesn't justify such a claim! Notwithstanding the appearances, corollary 4.2 puts in doubt the appropriateness of a formulation of an estimation problem where a constraint has been absorbed in the objective as a penalty term:

(i) Whereas one can immediately give some meaning to the choice of β in (CE) and find some justification for such a choice, the choice of any particular α in (PE) is much more difficult to justify since the interpretation must pass through corollary 4.2.

(ii) The choice of β in (CE) is independent the sample size whereas the choice of α in (PE) does depend on the sample size. Let's illustrate this in the context of Example 1.1, where Φ is the Fisher information, and the constraint $\Phi(h) \leq \beta$ has been included in the formulation of the problem to guarantee a certain level of smoothness for the estimated densities: $\hat{h}^1, \hat{h}^2, \dots, \hat{h}^\nu, \dots$. What corollary 4.2 tells us is that one could find $\alpha^1, \alpha^2, \dots, \alpha^\nu, \dots$ such that when this constraint is replaced by a penalty term $\alpha^\nu\Phi(h)$, $\nu = 1, 2, \dots$, the estimated densities obtained by solving the corresponding penalized estimation problem will have the required level of smoothness. The important observation here is that the coefficient to be assigned to this penalty term *depends* on the sample size ν , and thus a “good” coefficient for a given sample size might not be appropriate when the sample size is increased or decreased.

(iii) There are, at least conceptually, no new difficulties when introducing additional constraints in (CE) , whereas if more than one penalty term is included in (PE) , the choice of the appropriate coefficients for the additional penalty terms requires solving a problem that lies at the heart of constrained optimization.

Example 4.3. *The estimator of de Montricher, Tapia and Thompson.* The penalty term selected by de Montricher, Tapia and Thompson [9, 44] is simply

$$\alpha\Phi(h) = \alpha|h|, \quad \alpha > 0,$$

with

$$\begin{aligned} H &= H_0^s([\xi_l, \xi_u]) \\ &= \{ h \in L^2([\xi_l, \xi_u]; \mathbb{R}) \mid h^{(k)} \in L^2, h^{(k-1)}(\xi_l) = h^{(k-1)}(\xi_u) = 0, \quad k = 1, \dots, s \} \end{aligned}$$

with the $h^{(k)}$ distributional derivatives. Then Φ is a continuous convex function, and for any $\beta > 0$, one can find a density \tilde{h} such that $|\tilde{h}| < \beta$. In view of corollary 4.2, this means that the problem could equally well have been formulated as a constrained maximum likelihood estimation problem with the (single) constraint:

$$|h| \leq \beta$$

for some $\beta > 0$ (since the $|h| > 0$ for any density h). □

Example 4.4. *The first estimator of Good and Gaskins.* In [18], Good and Gaskins propose two specific estimators. The first one of these is obtained by solving (PE) with

$$\Phi(h) = \int_{\Xi} \frac{h'(\xi)^2}{h(\xi)} d\xi, \quad \text{where } \Xi = [\xi_l, \xi_u] \subset \mathbb{R},$$

and

$$H^\dagger = \{ h \mid h^{1/2} \in H^1([\xi_l, \xi_u]) \}.$$

H^\dagger is not a Hilbert space, not even a linear space, and Φ is not convex. So, corollary 4.2 doesn't apply, at least not directly. There is more than one way to pass from the resulting penalized estimation problem to an equivalent constrained problem. Let's start of as in [44, §4.3], since it leads to a convex optimization problem: First, because

$$\begin{aligned} \alpha \int [(h^{1/2})'(\xi)]^2 d\xi &= \frac{\alpha}{4} \int_{\Xi} \frac{h'(\xi)^2}{h(\xi)} d\xi, \\ -E^\nu \{ \ln h^{1/2}(\boldsymbol{\xi}) \} &= -\frac{1}{2} E^\nu \{ \ln h(\boldsymbol{\xi}) \}, \end{aligned}$$

one has,

$$\hat{h} \in \operatorname{argmin}_{h \in C^\dagger} \left[-E^\nu \{ \ln h(\boldsymbol{\xi}) \} + \alpha \int \frac{h'(\xi)^2}{h(\xi)} d\xi \right]$$

where

$$C^\dagger = \left\{ h \in H^\dagger \mid \int h(\xi) d\xi = 1, h \geq 0 \right\},$$

if and only if

$$\hat{g} := (\hat{h})^{1/2} \in \operatorname{argmin}_{g \in C} \left[-E^\nu \{ \ln g(\boldsymbol{\xi}) \} + 2\alpha \int g'(\xi)^2 d\xi \right]$$

where

$$C = \left\{ g \in H^1([\xi_l, \xi_u]) \mid \int g(\xi)^2 d\xi \leq 1, h \geq 0 \right\}.$$

The function \hat{g} is not a density function, but can be identified with a density that solves the originally formulated estimation problem; essentially, one can pass from one problem to the other by a change of variable and relaxing the constraint $\int g(\xi)^2 d\xi = 1$, which has no effect on the optimal solution(s). Next observe that

$$g \mapsto \int g'(\xi)^2 d\xi : H^1([\xi_l, \xi_u]) \rightarrow \mathbb{R},$$

is a convex function. Thus, assuming that $\alpha > 0$, in view of lemma 4.1, there exists β such that

$$\hat{g} \in \operatorname{argmin}_{g \in C} \left[-E^\nu \{ \ln g(\boldsymbol{\xi}) \} \mid \int g'(\xi)^2 d\xi \leq \beta \right].$$

The same change of variable $h^{1/2} = g$ leads to an equivalent constrained maximum likelihood density estimation problem. \square

Example 4.5. *The second estimator of Good and Gaskins.* The second estimator suggested by Good and Gaskins in [18] is

$$\hat{h} \in \operatorname{argmin}_{h \in C^\dagger} \left[-E^\nu \{ \ln h(\boldsymbol{\xi}) \} + \alpha \Phi(h^{1/2}) \right]$$

where

$$\Phi(h) = \int_{\Xi} h'(\xi)^2 d\xi + \frac{\gamma}{\alpha} \int_{\Xi} h''(\xi)^2 d\xi,$$

and

$$C^\dagger = \left\{ h \mid h^{1/2} \in H^2([\xi_l, \xi_u]), \int h(\xi) d\xi = 1, h \geq 0 \right\}.$$

Making the following change of variable, $g = h^{1/2}$, one is lead to the following “equivalent” convex optimization problem:

$$\min_{g \in C} \left[-E^\nu \{ \ln g(\boldsymbol{\xi}) \} + 2\alpha \Phi(g) \right]$$

where

$$C = \left\{ g \in H^2([\xi_l, \xi_u]) \mid \int g(\xi)^2 d\xi \leq 1, g \geq 0 \right\}.$$

The same arguments as those in example 4.4 allow the formulation of an equivalent constrained density estimation problem. \square

5. Epi-convergence: a primer

For the purposes of this paper, it will suffice to restrict our attention to functions defined on a separable Hilbert space $(H, |\cdot|)$. Let $f : H \rightarrow \overline{\mathbb{R}}$ be an extended real-valued function on H . Its *epigraph* is the set:

$$\text{epi } f := \{ (x, \alpha) \in H \times \mathbb{R} \mid f(x) \leq \alpha \}, \quad (5.1)$$

i.e., all the points in $H \times \mathbb{R}$ that lie on and above the graph of f . Observe that f is *lower semicontinuous (lsc)* if and only if $\text{epi } f$ is closed; recall that a function $f : H \rightarrow \overline{\mathbb{R}}$ is lower semicontinuous at x if $\liminf_{x' \rightarrow x} f(x') \geq f(x)$.

Definition 5.1. A sequence $\{f^\nu : H \rightarrow \overline{\mathbb{R}}, \nu \in \mathbb{N}\}$ *epi-converges* to $f : H \rightarrow \overline{\mathbb{R}}$ at x , if

$$\liminf_{\nu \rightarrow \infty} f^\nu(x^\nu) \geq f(x), \quad \forall x^\nu \rightarrow x; \quad (5.2)$$

and

$$\exists x^\nu \rightarrow x \text{ such that } \limsup_{\nu \rightarrow \infty} f^\nu(x^\nu) \leq f(x) \quad (5.3)$$

If this holds for all $x \in H$, the functions f^ν *epi-converge* to f , f is called the *epi-limit* of the f^ν , and one writes $f = \text{epi-lim}_{\nu \rightarrow \infty} f^\nu$ or $f^\nu \xrightarrow{e} f$. The name “*epi-convergence*” is motivated by the fact that this convergence notion is equivalent to the set-convergence of the epigraphs.

Epi-convergence yields the convergence of minimizers and optimal values, in a sense that will be made precise below, and it’s all that’s needed in many instances, in particular when H is finite dimensional. However, in infinite dimension, it turns out that it is useful to introduce a somewhat stronger notion, namely *Mosco-epi-convergence* which requires epi-convergence with respect to both the weak and the strong topologies.

Definition 5.2. A sequence $\{f^\nu : X \rightarrow \overline{\mathbb{R}}, \nu \in \mathbb{N}\}$, with $(H, |\cdot|)$ a Hilbert space, *Mosco-epi-converges* to $f : X \rightarrow \overline{\mathbb{R}}$ at x , if

$$\text{for all } x^\nu \xrightarrow{w} x \text{ (weak convergence), } \liminf_{\nu \rightarrow \infty} f^\nu(x^\nu) \geq f(x); \quad (5.4)$$

and

$$\exists x^\nu \rightarrow x \text{ (strong convergence) such that } \limsup_{\nu \rightarrow \infty} f^\nu(x) \leq f(x). \quad (5.5)$$

If this is the case for all $x \in X$, the functions f^ν *Mosco-epi-converge* to f , and one writes $f^\nu \xrightarrow{M:e} f$ or $f = M:\text{epi-lim}_{\nu \rightarrow \infty} f^\nu$.

These two definitions should, more precisely, be qualified as “*sequential*”, but it won’t be necessary to introduce this distinction here.

Theorem 5.3. Suppose $\{f, f^\nu : X \rightarrow \overline{\mathbb{R}}, \nu \in \mathbb{N}\}$ are such that $f^\nu \xrightarrow{e} f$, then

$$\limsup_{\nu \rightarrow \infty} (\inf f^\nu) \leq \inf f. \quad (5.6)$$

Moreover, if there is a subsequence $\{\nu_k\}_{k \in \mathbb{N}}$, such that for all k , $x^k \in \operatorname{argmin} f^{\nu_k}$ and $x^k \rightarrow \bar{x}$, then $\bar{x} \in \operatorname{argmin} f$ and also $\inf f^{\nu_k} \rightarrow \inf f$.

If the functions f^ν Mosco-epi-converges to f , and there is a subsequence $\{\nu_k\}_{k \in \mathbb{N}}$, such that for all k , $x^k \in \operatorname{argmin} f^{\nu_k}$ and $x^k \xrightarrow{w} \bar{x}$, then $\bar{x} \in \operatorname{argmin} f$ and $\inf f^{\nu_k} \rightarrow \inf f$.

Proof. These results are well known. We provide an elementary proof to illustrate the role played by the conditions (5.2), ..., (5.5). The inequality $\limsup_{\nu} \inf f^\nu \leq \inf f$ certainly holds if $\inf f = \infty$. If $\inf f$ is finite, then for all $\varepsilon > 0$ there exists x_ε such that $f(x_\varepsilon) < \inf f + \varepsilon$. Because the functions f^ν epi-converge to f , there exists a sequence $x^\nu \rightarrow x_\varepsilon$ such that $\limsup_{\nu} f^\nu(x^\nu) \leq f(x_\varepsilon) < \inf f + \varepsilon$. This implies that $\limsup_{\nu} \inf f^\nu < \inf f + \varepsilon$, and since this holds for all $\varepsilon > 0$, it yields the desired equality. The case when $\inf f = -\infty$ can be argued similarly, except that one now start with the observation that for all $\kappa > 0$ there exists x_κ such that $f(x_\kappa) < -\kappa$. Again (5.3) yields a sequence $x^\nu \rightarrow x_\kappa$ such that $\limsup_{\nu} f^\nu(x^\nu) \leq f(x_\kappa) < -\kappa$, which again implies that $\limsup_{\nu} \inf f^\nu < -\kappa$. Since this holds for all $\kappa > 0$, it follows that $\limsup_{\nu} \inf f^\nu = -\infty = \inf f$.

Now let $\{x^k, k \in \mathbb{N}\}$ be such that $x^k \in \operatorname{argmin} f^{\nu_k}$ for some subsequence $\{\nu_k\}_{k \in \mathbb{N}}$, and $x^k \rightarrow \bar{x}$. From (5.2), it follows,

$$\liminf_k (\inf f^{\nu_k}) = \liminf_k f^{\nu_k}(x^k) \geq f(\bar{x}).$$

On the other hand,

$$\inf f \geq \limsup_k (\inf f^{\nu_k}) \geq \liminf_k (\inf f^{\nu_k}),$$

with the first inequality following from the argument above. Hence, $f(\bar{x}) = \inf f$, i.e., $\bar{x} \in \operatorname{argmin} f$. Moreover, this implies that the inequalities in the two preceding identities are actually equalities, and consequently $\inf f^{\nu_k} \rightarrow \inf f$.

In the case of Mosco-epi-convergence, the argument is the same, except that the x^k converge weakly to \bar{x} and one appeals to (5.4) instead of (5.2). \square

There are other results of the epi-convergence theory that are important in a statistical setting, in particular those characterizing epi-convergence in terms of the convergence of the (sub)level sets, cf. [17].

6. Random lower semicontinuous functions

As already mentioned in §2, the objective functions $\mathbf{E}^\nu \mathbf{L}$ of the density estimation problems (2.2) are random functions since they depend on the empirical measure \mathbf{P}^ν that in turn depends on the samples collected for the random variables ξ_1, \dots, ξ_ν . This will be handled in the following framework: Let (Ξ, \mathcal{A}, P) be the (underlying) probability space with Ξ — the support of P — a Borel subset of \mathbb{R}^d , P the probability distribution of the \mathbb{R}^d -valued random variable ξ , and throughout it will be assumed that \mathcal{A} is P -complete. This last assumption can be dispensed with, but then the definition of a random lsc function, that is to follow, needs to be slightly modified, and some further technical development is then required that is better dispensed with at this stage.

It will be assumed throughout that H is a separable Hilbert space.

Definition 6.1. *A function $f : \Xi \times H \rightarrow \overline{\mathbb{R}}$ is a random lower semicontinuous (random lsc) function if*

- (i) *for all $\xi \in \Xi$, the function $x \mapsto f(\xi, x)$ is lower semicontinuous,*
- (ii) *$(\xi, x) \mapsto f(\xi, x)$ is $\mathcal{A} \otimes \mathcal{B}$ -measurable, where \mathcal{B} is the Borel field on H .*

The notion of a random lsc function, under the name “normal integrand,” was introduced by Rockafellar [33] in the context of the calculus of variations. In particular, he showed that in the present set-up, f is a random lsc function if and only if the set-valued mapping $\xi \mapsto \text{epi } f(\xi, \cdot)$ is a random closed set; recall that a set-valued mapping $\Gamma : \Xi \rightrightarrows H \times \mathbb{R}$ is a *random closed set* if $\Gamma(\xi)$ is closed for all $\xi \in \Xi$ and it is measurable, i.e., for all closed subsets $F \subset H \times \mathbb{R}$,

$$\Gamma^{-1}(F) := \{ \xi \in \Xi \mid \Gamma(\xi) \cap F \neq \emptyset \} \in \mathcal{A}.$$

Since H is separable, which implies that every open set can be written as the countable union of closed sets, it follows that also for every open set $G \subset H \times \mathbb{R}$,

$$\Gamma^{-1}(G) := \{ \xi \in \Xi \mid \Gamma(\xi) \cap G \neq \emptyset \} \in \mathcal{A}.$$

The following properties of random closed sets and random lsc functions are going to be needed.

Lemma 6.2. *For f^1 and f^2 be random lsc functions defined on $\Xi \times H$ and $\beta_1, \beta_2 \in \mathbb{R}_+$, the function $\beta_1 f^1 + \beta_2 f^2$ is another random lsc function.*

Proof. For all $\xi \mapsto \beta_1 f^1(\xi, \cdot) + \beta_2 f^2(\xi, \cdot)$ is lsc as follows immediately from the definition of lower semicontinuity. And $\beta_1 f^1 + \beta_2 f^2$ is $\mathcal{A} \otimes \mathcal{B}$ -measurable since measurability is preserved under linear combinations. \square

Proposition 6.3. *Let $f : \Xi \times H \rightarrow \overline{\mathbb{R}}$ be a random lsc function. Then, the infimal function*

$$\xi \mapsto \inf f(\xi, \cdot) := \inf_{x \in H} f(\xi, x) \text{ is } \mathcal{A}\text{-measurable,}$$

and the set of optimal solutions

$$\xi \mapsto \operatorname{argmin} f(\xi, \cdot) : \Xi \rightrightarrows H \text{ is a random closed set.}$$

Proof. Also these results are well known. The proof is based on the arguments in [34, theorems 3.45 and 3.47]. Let $\Gamma(\xi) = \operatorname{epi} f(\xi, \cdot)$, $\iota(\xi) = \inf f(\xi, \cdot)$, and $A(\xi) = \operatorname{argmin} f(\xi, \cdot)$. For $\beta \in \mathbb{R}$,

$$\iota^{-1}(-\infty, \beta) = \{\xi \in \Xi \mid \iota(\xi) < \beta\} = \Gamma^{-1}(H \times (-\infty, \beta)).$$

These sets belong to \mathcal{A} since Γ is a measurable random closed set and $H \times (-\infty, \beta)$ is an open subset of $H \times \mathbb{R}$.

Now observe that the function g defined by $g(\xi, x) = f(\xi, x) - \iota(\xi)$ is another random lsc function; using the convention $\infty - \infty = \infty$. Then,

$$\operatorname{argmin} f(\xi, \cdot) = \operatorname{lev}_0 g(\xi, \cdot) := \{x \in H \mid g(\xi, x) \leq 0\}.$$

The proof will be complete if one shows that for any $\alpha \in \mathbb{R}$, the level set mapping $\xi \mapsto \operatorname{lev}_\alpha f(\xi, \cdot) : \Xi \rightrightarrows H$ associated with the random lsc function f is a random closed set. With $\Lambda(\xi) = \operatorname{lev}_\alpha f(\xi, \cdot)$, for any closed set $F \subset H$,

$$\Lambda^{-1}(F) = \Gamma^{-1}(F \times \{\alpha\}) \in \mathcal{A},$$

since $F \times \{\alpha\}$ is a closed subset of $H \times \mathbb{R}$. Moreover, $\Lambda(\xi) = \{x \mid (x, \alpha) \in \operatorname{epi} f(\xi, \cdot)\}$ is closed for all ξ , i.e., Λ is a random closed set. \square

Proposition 6.4 [8, Lemma III.14]. *A function $f : \Xi \times H \rightarrow \overline{\mathbb{R}}$ that is \mathcal{A} -measurable in ξ and continuous in h is a random lsc function.*

Proof. It suffices to show that f is $\mathcal{A} \otimes \mathcal{B}$ -measurable. Let $\{h^\iota, \iota \in \mathbb{N}\}$ be a countable dense set in H . Set

$$f^\nu(\xi, h) := f(\xi, h^\iota) \text{ where } \iota = \inf_N [j \mid h \in \mathcal{B}^o(h^j, \nu^{-1})]$$

where $\mathcal{B}^o(h^j, \nu^{-1})$ is the open ball centered at h^j and radius ν^{-1} . Clearly, $f^\nu(\xi, h) \rightarrow f(\xi, h)$ as $\nu \rightarrow \infty$. Moreover, the functions f^ν are $\mathcal{A} \otimes \mathcal{B}$ -measurable, since

$$f^\nu(\xi, h) = f(\xi, h^\iota) \text{ on } \Xi \times \left[\mathcal{B}^o(h^\iota, \nu^{-1}) \setminus \bigcup_{j < \iota} \mathcal{B}^o(h^j, \nu^{-1}) \right],$$

and thus, also f is $\mathcal{A} \otimes \mathcal{B}$ -measurable. \square

For further properties of random lsc functions, consult [37, 46, 3] and the references therein.

Let $\text{SC}(H)$ be the *space of lower semicontinuous functions* defined on H . Again note that there is a one-to-one correspondence between the space of closed subsets of $H \times \mathbb{R}$ that are epigraphs and $\text{SC}(H)$; a set $E \subset H \times \mathbb{R}$ is an epigraph if $(x, \beta) \in E \Rightarrow E \cap (\{x\} \times \mathbb{R})$ is either \mathbb{R} or $[\tilde{\beta}, \infty)$ for some $\tilde{\beta} \leq \beta$. Let $\text{SC}_0(H)$ denote the space of *proper lsc functions*; a function g is *proper* if $g > -\infty$ and $g \neq \infty$.

The σ -field \mathcal{A}_f induced by the random lsc function f can be generated from the sets $\{\xi \in \Xi \mid \text{epi } f(\xi, \cdot) \cap G \neq \emptyset\}$ with G ranging over the open subsets of $H \times \mathbb{R}$, or equivalently by their complements, i.e., $\mathcal{A}_f = \sigma\text{-}\mathcal{E}$, where

$$\mathcal{E} = \left\{ \{\xi \in \Xi \mid \text{epi } f(\xi, \cdot) \cap G' \neq \emptyset\}, G' \subset H \times \mathbb{R} \text{ open} \right\};$$

one can even generate \mathcal{A}_f from a smaller collection \mathcal{E}' of sets, viz.,

$$\mathcal{A}_f = \sigma\text{-}\mathcal{E}' \quad \text{with } \mathcal{E}' = \left\{ \{\xi \in \Xi \mid \inf_G f > \alpha\} G \subset H \text{ open}, \alpha \in \mathbb{R} \right\}.$$

The *distribution* P_f of f refers to the probability measure induced on \mathcal{A}_f . Narrow (= weak) convergence of random lsc functions is defined in terms of the narrow convergence of their distributions. For more about convergence of random lsc functions, consult [37]. Random lower semicontinuous functions are *independent*, or *pairwise independent*, if the induced sigma-fields are independent, or pairwise independent. They are *identically distributed* if their distributions are identical. And, a family of random lsc functions is *iid* if they are independent and identically distributed, and *piid* if they are pairwise independent and identically distributed.

Lemma 6.5. *Let $f : \mathbb{R}^d \times H \rightarrow \overline{\mathbb{R}}$ a random lsc function, and $\xi_1, \xi_2 : (\Xi, \mathcal{A}) \rightarrow (\mathbb{R}^d, \mathcal{B})$ with $\Xi \subset \mathbb{R}^d$ iid random variables. Then, $\mathbf{f}^1, \mathbf{f}^2 : \Xi \times H \rightarrow \overline{\mathbb{R}}$ with $\mathbf{f}^\iota(\cdot, x) = f(\xi_\iota(\cdot), x)$ for $\iota = 1, 2$, are iid random lsc functions.*

Proof. With $\Gamma(\xi) = \text{epi } f(\xi, \cdot)$, and $\Gamma_\iota(\xi) = \text{epi } f^\iota(\xi, \cdot)$ for $\iota = 1, 2$, one has that for any open set $G' \subset H \times \mathbb{R}$,

$$\Gamma_\iota^{-1}(G') = \xi_\iota^{-1}(\Gamma^{-1}(G')) \quad \text{for } \iota = 1, 2.$$

Since f is a random lsc function, $\Gamma^{-1}(G') \in \mathcal{B}$, and consequently $\xi_\iota^{-1}(\Gamma^{-1}(G')) \in \mathcal{A}_\iota$ the σ -field induced by ξ^ι . Since \mathcal{A}_1 and \mathcal{A}_2 are independent, it follows from the construction above that the induced σ -fields \mathcal{A}_{f^1} and \mathcal{A}_{f^2} must also be independent. The preceding identities also imply that the random lsc functions $\mathbf{f}^1, \mathbf{f}^2$ will be identically distributed if ξ^1, ξ^2 are identically distributed. \square

7. Consistency

As already indicated in §2, consistency will be obtained by showing that the random functions $\mathbf{E}^\nu \mathbf{L}$ almost surely epi-converge to EL , or in other words, that the estimation problems are *epi-consistent* with a certain limit problem, identified in §3. The key to such an almost sure epi-convergence result is the law of large numbers for random lsc functions, see [25, 50, 3, 22, 1]. The particular versions that are going to be used here are recorded below.

In this section, to simplify notations, let $P = P^0$ denote the actual distribution of ξ . For a (measurable) function $\alpha : \Xi \rightarrow \overline{\mathbb{R}}$, its positive and negative parts are the functions $\alpha_+, \alpha_- : \Xi \rightarrow \overline{\mathbb{R}}_+$ given by $\alpha_+(\xi) = \max[0, \alpha(\xi)]$ and $\alpha_-(\xi) = \max[0, -\alpha(\xi)]$.

Theorem 7.1 [1, theorem 2.3]. *Let $(H, |\cdot|)$ be a separable Hilbert space, $\{\xi^\nu, \nu \in \mathbb{N}\}$ a sequence of piid (pairwise independent and identically distributed) random variables with common distribution P defined on (Ξ, \mathcal{A}) with \mathcal{A} P -complete, and let $f : \Xi \times H \rightarrow \overline{\mathbb{R}}$ be a random lsc function.*

Suppose that for all $x^0 \in H$ there is a neighborhood V of x^0 and a measurable function $\alpha : \Xi \rightarrow \mathbb{R}$ with $\int_{\Xi} \alpha_-(\xi) P(d\xi) < \infty$ such that P -almost surely $f(\cdot, x) \geq \alpha$ for all $x \in V$. Let \mathbf{P}^ν be the (random) empirical measure induced on (Ξ, \mathcal{A}) by the random variables ξ^1, \dots, ξ^ν , and let

$$\mathbf{E}^\nu \mathbf{f} := \int_{\Xi} f(\xi, \cdot) \mathbf{P}^\nu(d\xi)$$

be the (random) expectation of f with respect to \mathbf{P}^ν . Then, with $Ef(x) := \int_{\Xi} f(\xi, x) P(d\xi)$,

$$P^\infty\text{-almost surely: } \operatorname{epi}\text{-}\lim_{\nu \rightarrow \infty} \mathbf{E}^\nu \mathbf{f} = Ef.$$

Actually, the statement of this theorem is slightly more general than theorem 2.3 in [1]. It is assumed in [1] that the function α is summable, i.e., that $\int \alpha(\xi) P(d\xi)$ is finite. That's actually never used in the proof in [1], except when appealing to the classical version of Fatou's lemma to claim that Ef^1 is lsc. But that's also the case under this weaker assumption, cf. the Appendix. Moreover, by appealing to Etemadi's [15], rather than the standard, strong law of large numbers, one can relax the assumption of independent samples to only pairwise independence. The details are provided in the Appendix for f a random lsc function defined on $\Xi \times X$ with X a Polish space.

The assumption that the random lsc functions f^ν are minorized locally by a function α whose integral is bounded below is an essential component of the hypotheses. For purely technical reasons, this function will be assumed to be quadratic in theorem 7.2, this is predicated in part by the method of proof adopted in [3]. But, given the application that's going to be made of this theorem, this assumption is totally harmless. On the other hand, to obtain the almost sure Mosco-epi-convergence of $\mathbf{E}^\nu \mathbf{f}$ Ef it is going to be necessary to restrict ourselves to random *convex* lsc functions.

Theorem 7.2 [3, theorem 6.2]. Let $(H, |\cdot|)$ be a separable Hilbert space with $|\cdot|$ a smooth norm, $\{\xi^\nu, \nu \in \mathbb{N}\}$ a sequence of piid (pairwise independent and identically distributed) random variables with common distribution P defined on (Ξ, \mathcal{A}) with \mathcal{A} P -complete, and let $f : \Xi \times H \rightarrow \overline{\mathbb{R}}$ be a random convex lsc function. Suppose that for

$$P\text{-almost surely: } f(\xi, \cdot) \geq -\alpha_0 |\cdot - x^0|^2 + \alpha_1(\xi)$$

with $x^0 \in H$, $\alpha_0 \in \mathbb{R}_+$, $\alpha_1 \in \mathcal{L}^1((\Xi, \mathcal{A}, P); \mathbb{R})$, and

$$\exists v \in \mathcal{L}^2(\Xi; H) \text{ such that } \int_{\Xi} f(\xi, v(\xi)) P(d\xi) < \infty.$$

Let \mathbf{P}^ν be the (random) empirical measure induced on (Ξ, \mathcal{A}) by the random variables ξ^1, \dots, ξ^ν , and

$$\mathbf{E}^\nu \mathbf{f} := \int_{\Xi} f(\xi, \cdot) \mathbf{P}^\nu(d\xi)$$

be the (random) expectation of f with respect to \mathbf{P}^ν . Then, with $Ef(x) := \int_{\Xi} f(\xi, x) P(d\xi)$,

$$P^\infty\text{-almost surely: } \text{Mosco-epi-lim}_{\nu \rightarrow \infty} \mathbf{E}^\nu \mathbf{f} = Ef.$$

Proof. The statement of theorem 6.2 in [3] isn't quite in this form, it is slightly more general. It reads: Suppose $(X, |\cdot|)$ is a separable Banach space with $|\cdot|$ a smooth norm, $\{g^\nu : \Xi \times H \rightarrow \overline{\mathbb{R}}, \nu \in \mathbb{N}\}$ a sequence of piid random convex lsc functions P -almost surely bounded below by a quadratic minorant $g^\nu(\xi, x) \geq -\alpha_0 |x - x^0|^2 + \alpha_1(\xi)$ with $x^0 \in H$, $\alpha_0 \in \mathbb{R}_+$, $\alpha_1 \in \mathcal{L}^1((\Xi, \mathcal{A}, P); \mathbb{R})$ and suppose there exists $\hat{v} \in \mathcal{L}^2(\Xi; H)$ such that $\int g^1(\xi, \hat{v}(\xi)) P(d\xi) < \infty$. Then, P -almost surely

$$\frac{1}{\nu} \sum_{k=1}^{\nu} g^k(\xi, \cdot) \xrightarrow{M:e} Eg^1.$$

To pass from this to the statement of the theorem, one relies on lemma 6.5, and simply let $g^\nu(\cdot, x) = f(\xi^\nu(\cdot), x)$. \square

The consistency of the argmin estimator will follow from these laws of large numbers and theorem 5.3 (“convergence” of the argmin under epi-convergence) once it has been verified that

$$L(\xi, h) = \begin{cases} L_0(\xi, h) & \text{if } h \in S; \\ \infty & \text{otherwise;} \end{cases} \quad (7.1)$$

is a random lsc function, since

$$\mathbf{E}^\nu \mathbf{L}(h) = \frac{1}{\nu} \sum_{k=1}^{\nu} L(\xi^k, h); \quad (7.2)$$

recall that

$$L_0(\xi, h) = \begin{cases} -\ln h(\xi) & \text{if } h(\xi) > 0; \\ \infty & \text{otherwise.} \end{cases} \quad (7.3)$$

Let's begin with a couple of preparatory lemmas. Henceforth,

$$H \subset \mathcal{M} = \mathcal{M}(\Xi, \mathcal{B}_d; \mathbb{R}) = \{h : \Xi \rightarrow \mathbb{R} \mid h \text{ is Borel-measurable} \}$$

with \mathcal{B}_d is the Borel field relative to the Borel set $\Xi \subset \mathbb{R}^d$.

Lemma 7.3. *Let H be a Hilbert space. If for $\xi \in \Xi$, the valuation functional $h \mapsto h(\xi)$ is continuous on a set $C \subset H$, then the function $h \mapsto L_0(\xi, h)$ is continuous on C .*

Proof. For $\{h^\nu \in C, \nu \in \mathbb{N}\}$ a sequence converging to $h \in C$. One has to show that $\lim_\nu L_0(\xi, h^\nu) = L_0(\xi, h)$ for ξ an arbitrary point in Ξ . Continuity of the valuation functional on C means that $h^\nu(\xi) \rightarrow h(\xi)$ for all $\xi \in \Xi$ and all $h \in C$.

- (a) If $h(\xi) > 0$, then $h^\nu(\xi) > 0$ for ν sufficiently large, and this implies $L_0(\xi, h^\nu) = -\ln h^\nu(\xi) \rightarrow -\ln h(\xi) = L_0(\xi, h)$.
- (b) If $h(\xi) < 0$, then $h^\nu(\xi) < 0$ for ν sufficiently large, and consequently $L_0(\xi, h^\nu) = \infty \rightarrow \infty = L_0(\xi, h)$.
- (c) Finally, if $h(\xi) = 0$, then given any $\mu > 0$, one can find $\delta > 0$ such that for any $\theta \in (0, \delta)$, $-\ln \theta > \mu$, and since for ν sufficiently large, $h^\nu(\xi) \in [-\delta, \delta]$, it follows that then $L_0(\xi, h) \geq \mu$ which implies $L_0(\xi, h^\nu) \rightarrow \infty = L_0(\xi, h)$. \square

Let $|h|_1$ and $|h|_\infty$ denote the \mathcal{L}^1 and \mathcal{L}^∞ norms of a (measurable) function $h : \Xi \rightarrow \mathbb{R}$.

Lemma 7.4. *Let $C \subset H \subset \mathcal{M}$ where $(H, |\cdot|)$ is a Hilbert space. If there exist $\eta > 0, \kappa > 0$ such that for all $h \in C$, $|h| \geq \eta$ and $|h|_\infty \leq \kappa$, then for P -almost all $\xi \in \Xi$, the valuation functional $h \mapsto h(\xi)$ is continuous on C .*

Proof. Since for all $h \in C$, $|h| \geq \eta$, and for P -almost all $\xi \in \Xi$, $|h(\xi)| \leq \kappa$, one has that P -almost surely $|h(\xi)| \leq (\kappa/\eta)|h|$ which yields the continuity of the valuation functional $h \mapsto h(\xi)$ for P -almost all $\xi \in \Xi$. \square

Theorem 7.5. *Let $(H, |\cdot|)$, with $H \subset \mathcal{M}$, be a Hilbert space and suppose that*

$$S = \left\{ h \in A \subset H \mid \int_{\Xi} h(\xi) d\xi = 1, h \geq 0 \text{ } P\text{-a.s.} \right\}$$

is closed.

(a) *If H is a reproducing kernel Hilbert space, then $L : \Xi \times H \rightarrow \overline{\mathbb{R}}$, as defined by (7.1), is a random lsc function.*

(b) *If $h \in S$ implies $|h| \geq \varepsilon$ and $|h|_\infty \leq \kappa$ for some $\varepsilon > 0, \kappa > 0$, then there is a set Ξ_1 of P -measure 1 such that $L : \Xi_1 \times H$ is a random lsc function.*

Moreover, if A is convex, then L is a random convex lsc function.

Proof. If H is a reproducing kernel Hilbert space, the valuation functionals $h \mapsto h(\xi)$ are continuous. This is also the case, at least for all ξ in some set Ξ_1 of P -measure 1, when there are $\varepsilon, \kappa > 0$ such that $h \in S$ implies $|h| \geq \varepsilon$ and $|h|_\infty \leq \kappa$, cf. lemma 7.4. Thus, by lemma 7.3, $h \mapsto L_0(\xi, h)$ is continuous for all $\xi \in \Xi$ or all $\xi \in \Xi_1$, as the case might be. It then follows from proposition 6.4 that L_0 is a random lsc function. Since, S is $|\cdot|$ -closed, by assumption, the indicator function,

$$\delta_S(h) = \begin{cases} 0 & \text{if } h \in S, \\ \infty & \text{otherwise,} \end{cases}$$

is a random lsc function. It is lsc in h , and trivially jointly measurable. It now suffices to appeal to lemma 6.2 to complete the proof, since $L(\xi, h) = L_0(\xi, h) + \delta_S(h)$.

The convexity of $L(\xi, \cdot)$, when S is convex, follows from the convexity of δ_S and $L_0(\xi, \cdot)$ which, in turn, follows immediately from the concavity of \ln . \square

The pivotal role played by lemma 7.3 in this proof underscores the emphasis placed in statistical applications on reproducing kernel Hilbert spaces. Of particular interest here is the fact that when Ξ is an open set and $p = 1, 2, \dots$, the Sobolev spaces

$$\begin{aligned} H^p(\Xi) &= \{ \{ h \in \mathcal{M}(\Xi; \mathbb{R}) \mid h^{(n)} \in \mathcal{L}^2(\mathbb{R}^d; \mathbb{R}), n = 0, \dots, p \}, \\ H_0^p(\Xi) &= \{ h \in \mathcal{L}^2(\mathbb{R}^d; \mathbb{R}) \mid h^{(n)} \in \mathcal{L}^2(\mathbb{R}^d; \mathbb{R}), h^{(n-1)} = 0 \text{ on } \text{bdry } \Xi, n = 1, \dots, p \} \end{aligned}$$

are reproducing kernel Hilbert spaces; again derivatives are to be understood in the sense of distributions. Hence, when H is $H^p(\Xi)$ or $H_0^p(\Xi)$, in particular if

$$H = H^1(\Xi) = \{ h \in \mathcal{L}^2(\mathbb{R}^d; \mathbb{R}) \mid h' \in \mathcal{L}^2(\mathbb{R}^d; \mathbb{R}) \},$$

L_0 is a random lsc function on $\Xi \times H$. However, in such cases, there is no guarantee that S is closed, since the integration functional,

$$h \mapsto \int_{\Xi} h(\xi) d\xi,$$

is not, in general, continuous on $H^1(\Xi)$. This means that the set

$$\left\{ h \in H^1(\Xi) \mid \int_{\Xi} h(\xi) d\xi = 1 \right\}$$

is not necessarily closed. For a simple example, consider the (density) functions h^ν where

$$h^\nu(\xi) = \begin{cases} [\xi \ln \nu]^{-1} & \text{if } \xi \in (1, \nu), \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, for all ν , $\int_1^\infty h^\nu(\xi) d\xi = 1$, but $|h^\nu|_{H^1} \rightarrow 0$, i.e., $h^\nu \rightarrow h \equiv 0$ which certainly doesn't sum up to 1.

Remark 7.6. If $(H, |\cdot|)$ with $H \subset \mathcal{M}$, is a reproducing kernel Hilbert space, the valuation functionals $h \mapsto h(\xi)$ are continuous, and consequently $h^\nu(\xi) \rightarrow h(\xi)$ for all $\xi \in \Xi$ whenever $h^\nu |\cdot|$ -converges to h . In view of the dominated convergence theorem, to have $\int h^\nu(\xi) d\xi \rightarrow \int h(\xi) d\xi$, it would certainly suffice for $\xi \mapsto \sup_{h \in S} h(\xi)$ to be a summable function, a relatively mild restriction. \square

Remark 7.7. If $(H, |\cdot|)$, with $H \subset \mathcal{M}$, is a Hilbert space, and $h \in S$ implies $|h| \geq \varepsilon$, $|h|_\infty \leq \kappa$ for some $\varepsilon > 0$, $\kappa > 0$. Then in view of lemma 7.4, and again, the dominated

convergence theorem, one would have that $\int_{\Xi} h^\nu(\xi) d\xi \rightarrow \int_{\Xi} h(\xi) d\xi$ whenever $h^\nu \in S$ $|\cdot|$ -converge to h and Ξ was bounded. \square

Remark 7.8. Note that the condition $|h| \geq \varepsilon$ for some $\varepsilon > 0$ will be automatically satisfied in the following circumstances: Let $\Xi \subset \mathbb{R}^d$, and either $H = \mathcal{L}^2(\Xi; \mathbb{R})$ or $|h| \geq \gamma|h|_2$ for some $\gamma > 0$ with $|\cdot|_2$ the \mathcal{L}^2 -norm. Since $h \in S$ implies $h \geq 0$ and $\int_{\Xi} h(\xi) d\xi = 1$, given any $0 < \eta < 1$, one can find $\rho > 0$ such that

$$1 - \eta \leq \int_{\rho B} h(\xi) d\xi \leq \left(\int_{\rho B} 1^2 d\xi \right)^{\frac{1}{2}} \left(\int_{\rho B} h(\xi)^2 d\xi \right)^{\frac{1}{2}} \leq \rho^{d/2} |h|_2;$$

the second inequality coming from Hölder's inequality. Hence, $|h|_2 \geq \varepsilon := (1 - \eta)\rho^{-d/2}$. \square

Theorem 7.9 (epi-consistency). *Let $(H, |\cdot|)$, with $H \subset \mathcal{M}$, be a separable Hilbert space and $\{\xi^1, \xi^2, \dots\}$ a sequence of pairwise independent, identically distributed random variables. Suppose that*

$$S = \{ h \in A \subset H \mid \int_{\Xi} h(\xi) d\xi = 1, h \geq 0 \text{ P-a.s.} \}$$

is closed, and either one of the following conditions is satisfied: H is a reproducing kernel Hilbert space and Ξ is bounded, or $h \in S$ implies $|h| \geq \varepsilon$ and $|h|_\infty \leq \kappa$ for some $\varepsilon > 0$, $\kappa > 0$. Then, the random lsc functions $\mathbf{E}^\nu L$ epi-converge P^∞ -almost surely to EL , with $E^\nu L$ and EL as defined by (7.1)-(7.3), i.e., they are P^∞ -a.s. epi-consistent.

Proof. It suffices to verify that the assumptions of theorem 7.1 are satisfied. The hypotheses imply, via theorem 7.5, that there always is a set Ξ_1 a set of P -measure 1 such that $L : \Xi_1 \times H \rightarrow \overline{\mathbb{R}}$ is a random lsc function. So, we as well assume that L is defined on $\Xi_1 \times H$. The properties of the logarithmic function yield:

$$\alpha(\xi) := 1 - h(\xi) \leq -\ln h(\xi), \quad \forall \xi \in \Xi$$

with $\alpha_-(\xi) = 0$ unless $h(\xi) > 1$ in which case $\alpha_-(\xi) = h(\xi) - 1$. In both cases, Ξ is bounded or $|h|_\infty \leq \kappa$, $\int \alpha_-(\xi) P(d\xi)$ is bounded, and thus from theorem 7.1 it follows that

$$\text{epi-lim}_{\nu \rightarrow \infty} \mathbf{E}^\nu f = Ef, \quad P^\infty\text{-almost surely.}$$

The theorem refers to $E^\nu L : \Xi_1^\infty \times H \rightarrow \overline{\mathbb{R}}$ as random lsc functions. To see this simply observe that $L(\xi^l(\cdot), \cdot)$ can be identified with a extended real-valued function defined on $\Xi_1^\infty \times H$. These are random lsc functions, and $E^\nu L$ is a just finite sum of such functions (lemma 6.2). \square

Corollary 7.10 (consistency). *Let's assume that the hypotheses of theorem 7.9 are satisfied, $\hat{h}^\nu \in \text{argmin } E^\nu L$, and h^∞ is cluster point of the sequence $\{h^1, h^2, \dots\}$. Then P^∞ -almost surely, $h^\infty \in \text{argmin } EL$. Consequently, if $\text{argmin } EL$ is a singleton, and h^0 , the true*

density, has not been excluded by the addition of the “prior” information, i.e., $h^0 \in A$, then P^∞ -almost surely, the cluster point of the \hat{h}^ν must be h^0 . If $h^0 \notin A$, then h^∞ is a density in A that minimizes the Kullback-Leibler discrepancy from h^0 .

Proof. The assertions follow from theorem 5.3 since P^∞ -almost surely, the random functions $\mathbf{E}^\nu \mathbf{L}$ epi-converge to EL . From the discussion in §3, it follows that whenever h^0 belongs A it minimizes EL . If h^0 doesn't belong to A , again from the discussion in §3, it follows that the densities in $\operatorname{argmin} EL$ then minimize the Kullback-Leibler discrepancy $K(h, h^0) = - \int h^0(\xi) \ln (h(\xi)/h^0(\xi)) d\xi$ from h^0 . \square

Although, in general one can't exclude the possibility that the optimization problems, $\min E^\nu L$ and $\min EL$, have more than one solution, this can only occur under very special circumstances that would have to involve serious nonconvexities introduced by the additional constraints. This is a consequence of the following proposition

Proposition 7.11 (strict convexity). *The functions $E^\nu L_0$ and EL_0 are strictly convex on their effective domain, i.e., where they are less than ∞ . In particular, this means that if the set A is convex (and S is nonempty), the sets $\operatorname{argmin} E^\nu L$ and $\operatorname{argmin} EL$ are singletons, unless they are empty.*

Proof. For any density functions $h_0, h_1 \in H$, with $h_\lambda = (1 - \lambda)h_0 + \lambda h_1$ for $\lambda \in (0, 1)$, from the strict concavity of \ln , one has

$$\ln h_\lambda(\xi) > (1 - \lambda) \ln h_0(\xi) + \lambda \ln h_1(\xi), \quad \text{when either } h_0(\xi) > 0 \text{ and/or } h_1(\xi) > 0.$$

Thus, for all $\lambda \in (0, 1)$,

$$L_0(h_\lambda, \xi) < (1 - \lambda)L_0(h_0, \xi) + \lambda L_0(h_1, \xi) \quad \text{unless both } h_0(\xi) = h_1(\xi) = 0.$$

This strict inequality is preserved when integrating both sides with respect to P^ν or P , unless both h_0 and h_1 are 0 on a set of positive measure. In this latter case, h_0 and h_1 do not belong to the effective domain of $E^\nu L_0$ or/and EL_0 .

If the constraints determine a convex set S , the strictly convex functions $E^\nu L_0$ and EL_0 admit unique solutions whenever the set of minimizers is nonempty. \square

So far, the consistency results don't guarantee that given a sequence of argmin-estimators \hat{h}^ν they will actually have a cluster point h^∞ . This will, of course, be the case if the set S determined by the constraints is compact. But that's too strong a condition to expect it to be satisfied in most applications. Weak-compactness can more readily be achieved, either because it's built in the formulation of the problem, or by imposing an additional constraint, such as one requiring that the solutions of the estimation problem be bounded. In such a situation it is possible to also assert the existence of a weak-cluster h^∞ to the sequence of argmin-estimators \hat{h}^ν provided one can prove P^∞ -almost sure Mosco-epi-convergence of $\mathbf{E}^\nu \mathbf{L}$ to EL .

Theorem 7.12 (Mosco-epi-consistency). *Let $(H, |\cdot|)$, with $H \subset \mathcal{M}$, be a separable Hilbert space with $|\cdot|$ a smooth norm, and $\{\xi^1, \xi^2, \dots\}$ a sequence of pairwise independent, identically distributed random variables. Suppose that the set A determined by the “prior” information is convex, and that*

$$S = \{ h \in A \subset H \mid \int_{\Xi} h(\xi) d\xi = 1, h \geq 0 \text{ P-a.s.} \}$$

is closed, and either one of the following conditions is satisfied:

- (a) H is a reproducing kernel Hilbert space, the function $\xi \mapsto \sup_{h \in S} h(\xi)$ is summable, and $h_0^{1/2} \in H$ for some $h_0 \in S$ with $EL_0(h_0) < \infty$;
- (b) $h \in S$ implies $|h| \geq \varepsilon$ and $|h|_\infty \leq \kappa$ for some $\varepsilon > 0$, $\kappa > 0$ and for some $h_0 \in S$, $EL_0(h_0) < \infty$.

Then, the convex random lsc functions $\mathbf{E}^\nu L$ Mosco-epi-converge P^∞ -almost surely to EL with $E^\nu L$ and EL as defined by (7.1)-(7.3), i.e., they are P^∞ -a.s. Mosco-epi-consistent.

Proof. As in the proof of theorem 7.9, one relies on theorem 7.5 to conclude that L is a convex random lsc function, the convexity of S following directly from the convexity of A . The theorem now follows directly from theorem 7.2 since the assumptions imply that $L_0(\xi, h) \geq \inf_{h \in S} (1 - h(\xi)) =: \alpha_1(\xi)$ is summable, and thus, with $\alpha_0 = 0$, $L(\xi, h) \geq \alpha_0 |h| + \alpha_1(\xi)$. Moreover, in case (a), with $j(\cdot, \xi) \equiv h_0^{1/2}$, one has $j \in \mathcal{L}^2(\Xi; H)$ and since $h_0 \in S$,

$$\int L(\xi, j(\cdot, \xi)) P(d\xi) = \int L_0(\xi, h_0) P(d\xi) = - \int_{\Xi} \ln h_0(\xi) P(d\xi) < \infty$$

with the last inequality coming from the assumption $EL_0(h_0) < \infty$.

In case (b), when $h \in S$, $|h|_\infty \leq \infty$, and again with $j(\cdot, \xi) \equiv h_0$, one has $j \in \mathcal{L}^2(\Xi; H)$ and $\int L(\xi, j(\cdot, \xi)) P(d\xi) < \infty$. \square

Corollary 7.13 (consistency). *Let's assume that the hypotheses of theorem 7.12 are satisfied and that S is nonempty and bounded, $\hat{h}^\nu = \operatorname{argmin} E^\nu L$. Then, the sequence $\{\hat{h}^\nu, \nu \in \mathbb{N}\}$ admits a weak-cluster point h^∞ that P^∞ -almost surely is the unique solution of $\min EL$. If h^0 , the true density, belongs to A , then P^∞ -almost surely $h^\infty = h^0$. If $h^0 \notin A$, then h^∞ is P^∞ -almost surely the density in A that minimizes the Kullback-Leibler discrepancy from h^0 .*

Proof. After observing that S is weakly compact, the assertions follows directly from the preceding theorem, theorem 5.3, the discussion in §3, and proposition 7.11 which allows us to claim the uniqueness of the solutions. \square

8. Implementation: an example

So far, the attention has been concentrated on the theoretical foundations of our approach. In this section, a simple example illustrates the overall strategy that will be used in the numerical implementation. A projected paper [11] will provide the details, and will report on the results obtained when estimating the densities of univariate *and* multivariate random variables.

Let $\xi^1, \xi^2, \dots, \xi^\nu$ be iid samples coming from a random variable ξ that has exponential distribution with $E\xi = 1$, i.e.,

$$h^0(\xi) = \begin{cases} e^{-\xi} & \text{if } \xi \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

To obtain the argmin-estimator h^ν , see (1.1), one has to solve the following infinite dimensional optimization problem:

$$\begin{aligned} \min \quad & - \sum_{\ell=1}^{\nu} \ln h(\xi^\ell) \\ \text{so that} \quad & \int_{\mathcal{R}} h(\xi) d\xi = 1, \\ & h(\xi) \geq 0, \forall \xi \in \mathcal{R}, \\ & h \in A \subset H. \end{aligned}$$

In general, such problems don't have a closed form solution, and one has to resort to numerical procedures to find a solution, and consequently, finite dimensional approximations schemes will have to be examined.

One could rely on a piecewise line approximation, implicitly implying that the density function is of class C^0 (zero order smoothness), see [44], for example. By relying on splines, say B-splines, one could guarantee a density estimator with a higher order of smoothness, see again [44]. But, there are some drawbacks to such an approach. One difficulty is that the set of knots required to anchor the splines is hard to choose. Another one, that would haven't played a role here but is important when dealing with the densities of multivariate random variables, is that there is no "practical" high dimensional spline theory that can be used in this context.

An alternative approach is to use a finite number of terms of an orthonormal basis for H . Let

$$\{\varphi^k : \Xi \rightarrow \mathcal{R}, k = 1, \dots\} \quad \text{with} \quad H = \text{linear-span} \{\varphi^1, \varphi^2, \dots\}$$

be such a basis. And now, replace the optimization problem by

$$\begin{aligned} \min \quad & - \sum_{\ell=1}^{\nu} \ln \sum_{k=1}^q u_k \varphi^k(\xi^\ell) \\ \text{so that} \quad & \sum_{k=1}^q u_k \int_{\mathbf{R}} \varphi^k(\xi) d\xi = 1, \\ & \sum_{k=1}^q u_k \varphi^k(\xi) \geq 0, \quad \forall \xi \in \mathbf{R}, \\ & \sum_{k=1}^q u_k \varphi^k \in A, \\ & u_k \in \mathbf{R}, \quad k = 1, \dots \end{aligned}$$

The optimization problem involves now a finite number of variables, viz. u_1, u_2, \dots, u_q , but there are still an infinite number of constraints, in particular the nonnegativity constraints. This is a so-called semi-infinite programming problem, cf. [23]. Algorithmic procedures to solve such problems come in various flavors depending on the structure of the problem at hand. One that we have used in certain cases is the Phase I-Phase II procedure of Polak and He [30]. Questions related to the choice of the approximating problem and the accompanying solution method will be addresses in much more detail in [11], as mentioned earlier.

Here, let's consider $H = \mathcal{L}^2([0, \theta]; \mathbf{R})$ with orthonormal base

$$\sqrt{1/\theta}, \quad \sqrt{2/\theta} \cos\left(\frac{k\pi\xi}{\theta}\right), \quad k = 1, 2, \dots,$$

and let

$$h_q(u, \xi) = u_0 \sqrt{1/\theta} + \sum_{k=1}^q u_k \sqrt{2/\theta} \cos\left(\frac{k\pi\xi}{\theta}\right).$$

The nonlinear semi-infinite programming problem becomes:

$$\begin{aligned} \min \quad & -\frac{1}{\nu} \sum_{\ell=1}^{\nu} \ln \left[\frac{1}{\sqrt{\theta}} u_0 + \frac{\sqrt{2}}{\sqrt{\theta}} \sum_{k=1}^q \cos\left(\frac{k\pi\xi^\ell}{\theta}\right) u_k \right] \\ \text{so that} \quad & u_0 + \sqrt{2} \sum_{k=1}^q \frac{\sin(k\pi)}{k\pi} u_k = 1/\sqrt{\theta}, \\ & u_0 + \sqrt{2} \sum_{k=1}^q \cos\left(\frac{k\pi\xi}{\theta}\right) u_k \geq 0, \quad \forall \xi \in [0, \theta], \\ & u_k \in \mathbf{R}, \quad k = 0, \dots, q, \end{aligned} \tag{UE}$$

when there is no additional information available about the random variable ξ .

If for example it is known that the density is a decreasing, more precisely nonincreasing, function on $[0, \theta]$, then we can add the constraints: $h_q(\xi) \geq h_q(\xi')$ whenever $\xi \leq \xi'$, and the estimation problem becomes:

$$\begin{aligned} \min \quad & -\frac{1}{\nu} \sum_{\ell=1}^{\nu} \ln \left[\frac{1}{\sqrt{\theta}} u_0 + \frac{\sqrt{2}}{\sqrt{\theta}} \sum_{k=1}^q \cos \left(\frac{k\pi\xi^\ell}{\theta} \right) u_k \right] \\ \text{so that} \quad & u_0 + \sqrt{2} \sum_{k=1}^q \frac{\sin(k\pi)}{k\pi} u_k = 1/\sqrt{\theta}, \\ & u_0 + \sqrt{2} \sum_{k=1}^q \cos \left(\frac{k\pi\xi}{\theta} \right) u_k \geq 0, \quad \forall \xi \in \mathbb{R}, \\ & \sum_{k=1}^q \left[\cos \left(\frac{k\pi\xi}{\theta} \right) - \cos \left(\frac{k\pi\xi'}{\theta} \right) \right] u_k \geq 0, \quad \forall 0 \leq \xi \leq \xi' \leq \theta, \\ & u_k \in \mathbb{R}, \quad k = 0, \dots, q. \end{aligned} \tag{CE}$$

The solutions to these two optimization problems, whose graphs appear in Figures 8–1 and 8–2, were obtained by replacing the infinite number of constraints by a finite number of them. The constraints $h_q(\xi_j) \geq 0$ for $0 \leq \xi_1 \leq \dots \leq \xi_J \leq \theta$, with J finite, were substituted for the infinite number of constraints $h_q(\xi) \geq 0$ for all $\xi \in [0, \theta]$, and a similar substitution was made for the constraints enforcing the solution to (CE) to be nonincreasing. The resulting problems are then finite dimensional optimization problems that can be solved by a number of available packages for nonlinear programming problems. We relied on a software package developed by, and graciously put at our disposal, by André Tits [27] that implements their version of the sequential quadratic programming method. It turned out that in both cases, the solutions obtained did not only satisfy the constraints at $\xi = \xi_1, \dots, \xi_J$, but also at all points $\xi \in [0, \theta]$. If that had not been the case, one could add to the collection $\{\xi_1, \dots, \xi_J\}$ a number of points to this “discretization” of $[0, \theta]$ by picking points ξ at which the proposed solution fails most significantly the nonnegativity and/or the nonincreasing condition.

The example to be dealt with numerically, has sample size $\nu = 20$. It is on purpose that ν has been selected quite small. Any reasonable nonparametric estimation method should work relatively well when the sample size is relatively large, but might fail to come up with believable results when the sample size is small. Kernel estimation techniques, for example, perform very poorly when ν is small.

The solution of (EU) is graphed Figure 8–1. The argmin-estimator was computed with $\theta = 4.1$ (substantially larger than any of the samples ξ^1, \dots, ξ^{20}) and $q = 3$, i.e., with four base functions; the mean square error was 0.02339 (20 replications). The use of a richer approximating basis, i.e., with $q > 3$ yields estimated densities that oscillate in the tail and have larger mean square errors.

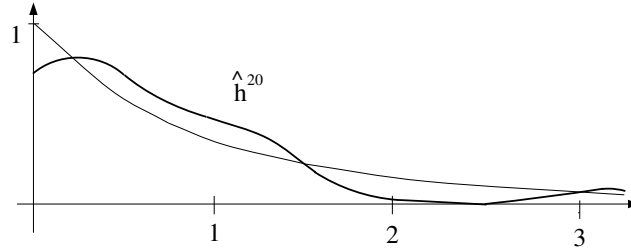


Fig. 8–1. argmin-estimator given 20 samples of an exponentially distributed random variable.

The solution of (EC) , that insists on densities that are nonincreasing, is graphed in Figure 8–2. With $\theta = 4.1$, and $q = 3$, one obtains a solution with mean square error below 0.02876, but an even better solution is obtained with $q = 5$, in which case the mean square error is just 0.008743 (again with 20 replications).

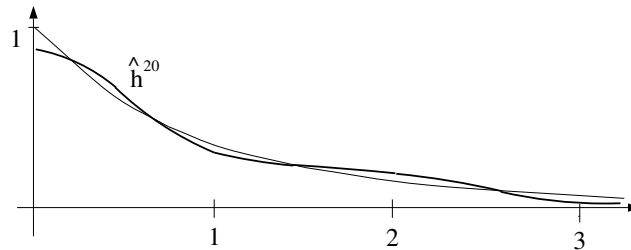


Fig. 8–2. argmin-estimator with monotonicity constraint.

From an information viewpoint, there is an explanation for this. Theoretically, it might appear as if an approximation involving a richer basis should always provide a more accurate approximation. But since the sample is quite small, or equivalently very limited information is available, the solution procedure can only reliably calculate a limited number of determining parameters. The inclusion of more base functions would imply that enough information is available to justify the determination of additional parameters. Hence, depending on the amount of information that is available, one can only estimate the density to a certain degree of accuracy and more terms do not necessarily generate better solutions, as is the case when solving (EU) with more than four base functions. When more information becomes available, for example, if it is known that the density function is monotone nonincreasing, then the density can be estimated to a higher degree accuracy when a richer basis is used in the formulation of (CE) . And this is actually the case here since more parameters can be calculated reliably, and the argmin-estimator gets significantly more accurate when solving (CE) with $q = 5$.

The relationship between the formulation of the approximating problems and the information available certainly deserves further exploration.

A. Appendix

We provide the proof of theorem 7.1. Essentially we reproduce that of [3, theorem 2.3] but with two (slight) modifications. The first one is the relaxation of the assumption that the integral of the lower bounding function α is finite. We now require only that the integral of its negative part be bounded, i.e.,

$$\int \alpha_-(\xi) P(d\xi) < \infty \quad \text{where } \alpha_-(\xi) = \max[0, -\alpha(\xi)].$$

And the second one is replacing the iid assumption with the somewhat weaker assumption that the random samples are *piid*, i.e., pairwise independent and identically distributed.

Let (Ξ, \mathcal{A}, P) be the underlying probability space with \mathcal{A} complete with respect to P , and (X, d) a Polish space, i.e., a complete separable metric space. The integral of a measurable extended real-valued function α is defined as $\int \alpha(\xi) P(d\xi) := \int \alpha_+(\xi) P(d\xi) - \int \alpha_-(\xi) P(d\xi)$ with α_+, α_- the positive and negative parts of the function α . The value to assign to $\int \alpha(\xi) P(d\xi)$ is clearly determined as long as both integrals of the nonnegative functions α_+ and α_- are not ∞ , in which case one could convene to set $\int \alpha(\xi) P(d\xi) = \infty$; this is consistent with the “minimization” context we have adopted here.

The following are the basic assumptions under which we are going to operate:

Assumption A.1. The function $f : \Xi \times X \rightarrow \overline{\mathbb{R}}$ is $\mathcal{A} \otimes \mathcal{B}$ -measurable where \mathcal{B} is the Borel field on X , and $f(\xi, \cdot)$ is lsc for all $\xi \in \Xi$, i.e., f is a *random lsc function*.

Assumption A.2. For each $x^0 \in X$ there exists a neighborhood V of x^0 and a measurable function $\alpha(\xi) : \Xi \rightarrow \overline{\mathbb{R}}$ such that $\int \alpha_-(\xi) P(d\xi) < \infty$ and

$$\text{for all } x \in V: \quad f(x, \xi) \geq \alpha(\xi).$$

Assumption A.3. The random variables ξ^1, ξ^2, \dots are pairwise independent and identically distributed.

Let’s begin with two preliminary lemmas. For Fatou’s lemma one could also consult [21].

Fatou’s Lemma (trivially generalized). *Let $\{\varphi^\nu : \Xi \rightarrow \overline{\mathbb{R}}, \nu \in \mathbb{N}\}$ be \mathcal{A} -measurable functions such that for all $\nu \in \mathbb{N}$, $\varphi^\nu \geq \alpha$ with $\alpha : \Xi \rightarrow \overline{\mathbb{R}}$ a \mathcal{A} -measurable function such that $\int \alpha_-(\xi) P(d\xi) < \infty$. Then*

$$\int_{\Xi} \liminf_{\nu \rightarrow \infty} \varphi^\nu(\xi) P(d\xi) \leq \liminf_{\nu \rightarrow \infty} \int_{\Xi} \varphi^\nu(\xi) P(d\xi).$$

Proof. First observe that $\varphi := \liminf_{\nu} \varphi^\nu$ is \mathcal{A} -measurable and that $\varphi \geq \alpha$. Since $\varphi \leq \varphi^\nu$ for all ν , and the integral is order preserving, one has

$$\int_{\Xi} \alpha(\xi) P(d\xi) \leq \int_{\Xi} \varphi(\xi) P(d\xi) \leq \int_{\Xi} \varphi^\nu(\xi) P(d\xi),$$

from which the assertion follows directly. \square

Beer's Construct [5]. Let $f : X \rightarrow \overline{\mathbb{R}}$ be a proper, lsc function that majorizes a real-valued function g .

The following construction:

$$\begin{aligned} f^1 &= g, \\ f^{\nu+1}(x) &= f^\nu(x) + d^\dagger((x, f^\nu(x)), \text{epi } f), \quad \nu \in \mathbb{N}, \end{aligned}$$

where

$$d^\dagger((x, \alpha), (x', \alpha')) := \max [d(x, x'), |\alpha - \alpha'|],$$

generates a sequence of functions $\{f^\nu : X \rightarrow \mathbb{R}, \nu \in \mathbb{N}\}$ with the following properties:

- (a) $g \leq f^\nu \leq f^{\nu+1} \leq f$ for all ν ;
- (b) $f^\nu(x) \rightarrow f(x)$ for all $x \in X$;
- (c) if g is locally lipschitzian, so are all the f^ν , and if g is lipschitzian, then so also are all the f^ν .

Proof. It is evident that the functions f^ν are real-valued, and for all x , the sequence $\{f^\nu(x)\}_{\nu \in \mathbb{N}}$ is monotone nondecreasing with $g(x)$ the lower bound and $f(x)$ and upper bound. That takes care of (a). If $f^\nu(x) \not\rightarrow f(x)$, then

$$g(x) + \sum_{\nu=1}^{\infty} d^\dagger((x, f^\nu(x)), \text{epi } f) < f(x)$$

which implies $\sum_{\nu=1}^{\infty} \max [d(x, x^\nu), |f^\nu(x) - \alpha^\nu|] < \infty$ where (x^ν, α^ν) is a point of $\text{epi } f$ that nearly minimizes, say up to $\varepsilon^\nu > 0$ with $\varepsilon^\nu \downarrow 0$, the distance between $(x, f^\nu(x))$ and $\text{epi } f$; remember that f proper guarantees $\text{epi } f$ nonempty. Thus, $x^\nu \rightarrow x$, $|f^\nu(x) - \alpha^\nu| \rightarrow 0$, and since f is lsc, one has $f(x) \leq \liminf_\nu f(x^\nu) = \liminf_\nu \alpha^\nu = \lim_\nu f^\nu(x) < f(x)$. Evidently a contradiction, and consequently $f^\nu(x) \rightarrow f(x)$.

Let $\text{lip } G$ or $\text{lip } f$ be the (smallest) lipschitzian constant that can be associated with a mapping G or a function f . For any mapping $G : X \rightarrow X \times \mathbb{R}$, one has $|d^\dagger(G(x_1), \text{epi } f) - d^\dagger(G(x_2), \text{epi } f)| \leq |G(x_1) - G(x_2)|$. If G is (locally) lipschitzian, this property will be inherited by $d^\dagger(\cdot, \text{epi } f) \circ G$. Now, if f^ν is (locally) lipschitzian, then $G^\nu(x) = (x, f^\nu(x))$ is (locally) lipschitzian with $\text{lip } G^\nu \leq \text{lip } f^\nu + 1$, and so is $d^\dagger(\cdot, \text{epi } f) \circ G^\nu$ with the same lipschitzian constant. Since $f^{\nu+1} = f^\nu + d^\dagger(\cdot, \text{epi } f) \circ G^\nu$, $f^{\nu+1}$ is (locally) lipschitzian with $\text{lip } f^{\nu+1} \leq 2 \text{lip } f^\nu + 1$. By induction it now follows that if $f^1 = g$ is (locally) lipschitzian, so are all f^ν . \square

Theorem A.4. Let (X, d) be a Polish space, $\{\xi^\nu, \nu \in \mathbb{N}\}$ a sequence of piid (pairwise independent and identically distributed) random variables with common distribution P defined on (Ξ, \mathcal{A}) with \mathcal{A} P -complete, and let $f : \Xi \times X \rightarrow \overline{\mathbb{R}}$ be a random lsc function.

Suppose that for all $x^0 \in X$ there is a neighborhood V of x^0 and a measurable function $\alpha : \Xi \rightarrow \mathbb{R}$ with $\int_{\Xi} \alpha_-(\xi) P(d\xi) < \infty$ such that P -almost surely $f(\cdot, x) \geq \alpha$ for all $x \in V$ and all ν . Let \mathbf{P}^ν be the (random) empirical measure induced on (Ξ, \mathcal{A}) by the random variables ξ^1, \dots, ξ^ν , and let

$$\mathbf{E}^\nu \mathbf{f} := \int_{\Xi} f(\xi, \cdot) \mathbf{P}^\nu(d\xi)$$

be the (random) expectation of f with respect to \mathbf{P}^ν . Then, with $Ef(x) = \int_{\Xi} f(\xi, x) P(d\xi)$,

$$\text{epi-lim}_{\nu \rightarrow \infty} \mathbf{E}^\nu \mathbf{f} = Ef, \quad P^\infty\text{-almost surely .}$$

Proof. Let's begin by observing that under the assumptions, that (the slightly generalized) Fatou's lemma implies that Ef is lsc. To prove epi-convergence, one has to show that P^∞ -almost surely condition (5.2), i.e.,

$$\text{for almost all } \xi^1, \xi^2, \dots : \liminf_{\nu \rightarrow \infty} \mathbf{E}^\nu \mathbf{f}(x^\nu) \geq Ef(x), \quad \forall x^\nu \rightarrow x, \quad (\text{A.1})$$

and P^∞ -almost surely condition (5.3), i.e.,

$$\text{for almost all } \xi^1, \xi^2, \dots : \exists x^\nu \rightarrow x : \limsup_{\nu \rightarrow \infty} \mathbf{E}^\nu \mathbf{f}(x^\nu) \leq Ef(x), \quad (\text{A.2})$$

are satisfied. Let's begin with the first one of these.

Fix $x^0 \in X$. Let V be an open neighborhood x^0 and $\alpha : \Xi \rightarrow \mathbb{R}$ the associated function such that $\int_{\Xi} \alpha_-(\xi) P(d\xi) < \infty$. We verify first that almost surely (A.1) holds for the restriction of $\mathbf{E}^\nu \mathbf{f}$ and of Ef to V .

One possibility is that on a subset of Ξ of positive measure, the function $f(\cdot, x) = \infty$ identically, for all $x \in V$. Then the result is trivial since then $Ef = \infty$ on V , and clearly, for all $x \in V$, P^∞ -almost surely $\mathbf{E}^\nu \mathbf{f}(x) = \infty$ when ν is large enough.

So, let's proceed under the assumption that P -almost surely $f(\xi, x) < \infty$ for at least some $x \in V$, and let's appeal to Beer's construction to generate a sequence of functions $\{g^k(\xi, x) : V \rightarrow \mathbb{R}, k \in \mathbb{N}\}$ with the following properties:

- (i) each g^k is a random lsc function on $\Xi \times V$;
- (ii) each g^k is lipschitzian in x , with lipschitzian constant independent of ξ ;
- (iii) $g^k(\xi, x) \geq \alpha(\xi)$, and $g^k(\xi, x)$ converges monotonically to $f^\nu(\xi, x)$ as $k \rightarrow \infty$.

Simply let,

$$\begin{aligned} g^0(\xi, x) &= \alpha(\xi), \\ g^{k+1}(\xi, x) &= g^k(\xi, x) + d^\dagger((x, g^k(\xi, x)), \text{epi } f(\xi, \cdot) \cap (V \times \mathbb{R})), \quad k = 0, 1, \dots, \end{aligned}$$

where $d^\dagger((x, \alpha), (x', \alpha')) = \max[d(x, x'), |\alpha - \alpha'|]$; note that $f(\xi, \cdot)$ is proper on V .

The lower semicontinuity of $g^k(\xi, \cdot)$ is immediate from the recursive definition, in fact it's lipschitzian as follows from Beer's construction. The measurability of g^k follows, by induction, from that of the mapping

$$(\xi, x) \mapsto \inf_{(z, \beta)} \left\{ \max [d(x, z), |\gamma(\xi, x) - \beta|] \mid \beta \geq f(\xi, z) \right\}$$

where γ is a random lsc function with $\gamma(\xi, \cdot) \leq f(\xi, \cdot)$, or equivalently, from that of

$$(\xi, x) \mapsto \inf_{z \in X} h(\xi, x; z) := \max [d(x, z), f(\xi, z) - \gamma(\xi, x)].$$

And, this is an immediate consequence of [32] after observing the following:

- 1) $h : (\Xi \times X) \times X$ is $(\mathcal{A} \otimes \mathcal{B}) \otimes \mathcal{B}$ -measurable;
- 2) for all $(\xi, x) \in \Xi \times X$, $h(\xi, x; \cdot)$ is lsc.

This takes care of property (i) claimed for g^k , while properties (ii) and (iii) are direct consequences of Beer's construction.

With the aid of the sequence $g^k(\xi, x)$, one can verify (A.1) on V as follows. For $x \in V$ and k fixed, define

$$\mathbf{E}^\nu g^k(x) := \int_{\Xi} g^k(\xi, x) P^\nu(d\xi)$$

which is bounded below since $g^k(\xi, x) \geq \alpha(\xi)$. The Etemadi's version of the strong law of large numbers (for real-valued random variables) [15] implies

$$\forall x \in V : \mathbf{E}^\nu g^k(x) \rightarrow E g^k(x) = \int_{\Xi} g^k(\xi, x) P(d\xi), \quad P^\infty\text{-almost surely.}$$

In fact, for $D = \{x^i, i = 1, \dots\}$ a countable dense subset of V , one has that P^∞ -almost surely, $\mathbf{E}^\nu g^k \rightarrow E g^k$ on D ; one exploits here the fact that D is countable. Since D is dense in V , the lipschitzian property in (ii) allows us to extend this to: $\mathbf{E}^\nu g^k \rightarrow E g^k$ on V P^∞ -almost surely. Finally, since the collection of functions $\{g^k, k \in \mathbb{N}\}$ is countable, one is able to conclude that except possibly for a set of null P^∞ -measure,

$$\text{for } k = 1, \dots, \forall x \in V : \mathbf{E}^\nu g^k(x) \rightarrow E g^k(x).$$

Next, let x be an arbitrary point in V , and suppose first that $E f(x) < \infty$. Since for any k , the lipschitzian constant of $g^k(\xi, \cdot)$ is shared by all the ξ (it is $2^k - 1$), hence $E g^k$ and for all ν , $\mathbf{E}^\nu g^k$ are also lipschitzian. Thus, except possibly for a set of P^∞ -measure 0, for any sequence $x^\nu \rightarrow x$ (in V):

$$\text{for all } k = 1, \dots, : \mathbf{E}^\nu g^k(x^\nu) \rightarrow E g^k(x).$$

On one hand, the monotone convergence property in (iii) implies that for k sufficiently large, $E g^k(x)$ is close to $E f(x)$, say $E f(x) - E g^k(x) < \varepsilon$ for some (small) $\varepsilon > 0$, and on the other hand, for all k , $f(\xi, \cdot) \geq g^k(\xi, \cdot)$. For k sufficiently large, this yields,

$$\liminf_{\nu \rightarrow \infty} \mathbf{E}^\nu f(x^\nu) \geq \lim_{\nu \rightarrow \infty} \mathbf{E}^\nu g^k(x^\nu) = E g^k(x) \geq E f(x) - \varepsilon \quad P^\infty\text{-a.s. .}$$

Since ε can be chosen arbitrarily small (adjusting k in the process), it follows that (A.1) is satisfied when $Ef(x) < \infty$. The case $(Ef)(x) = \infty$ is similar; the only modification is to replace the phrase ε arbitrarily small by $Eg^k(x)$ arbitrarily large.

So far, it has been verified that P^∞ -almost surely (A.1) holds for $x \in V$. Since X is assumed separable, a countable number of such neighborhoods V cover X . Therefore, there is a set of full P^∞ -measure on which the above holds for all $x \in X$.

We now verify that P^∞ -almost surely condition (A.2) for epi-convergence is satisfied. Consider the set $\text{epi}_+ Ef = \{(x, \beta) \in X \times \overline{\mathbb{R}} \mid \beta \geq Ef(x)\}$ as a (nonempty) subset of $X \times \overline{\mathbb{R}}$, and let $D_+ = \{(x^i, \beta_i), i \in \mathbb{N}\}$ be a dense countable subset of $\text{epi}_+ Ef$; note that $\beta_i = \infty$ is allowed. The lower semicontinuity of Ef implies in particular that the set $D := \{(x^i, Ef(x^i)), i \in \mathbb{N}\}$ is dense in the lower boundary of the $\text{epi}_+ Ef$. Namely, for each $x \in X$, there is a subsequence of points in D that converges to $(x, Ef(x))$. Applying again Etemadi's version of the strong law of large numbers for each x^i separately, and using the countability of D , one concludes that on a set of full P^∞ -measure, $\mathbf{E}^\nu \mathbf{f}(x^i)$ converges to $Ef(x^i)$ for all x^i with $(x^i, Ef(x^i)) \in D$.

Clearly, for any given x one can find $\{x^\nu, \nu \in \mathbb{N}\}$ with $(x^\nu, Ef(x^\nu)) \in D$ and $\mathbf{E}^\nu \mathbf{f}(x^\nu) \rightarrow Ef(x)$ P^∞ -almost surely, and this then verifies condition (A.2). \square

Acknowledgement. We are grateful for the discussions we have had on nonparametric estimation, and statistical estimation in general, with Georg Pflug (University of Vienna) and Charles Geyer (University of Minnesota). The presentation was also helped by the useful comments received from Jitka Dupačová and Petr Lachout (Charles University, Prague).

References

- [1] Zvi Artstein & Roger J-B Wets, "Consistency of minimizers and the SLLN for stochastic programs," *J. of Convex Analysis* 1 (1995), 1–17.
- [2] Hedy Attouch, *Variational Convergence for Functions and Operators*, Pitman, London, 1984, Applicable Mathematics Series.
- [3] Hedy Attouch & Roger J-B Wets, "Epigraphical processes: laws of large numbers for random lsc functions," *Séminaire d'Analyse Convexe, Montpellier* (1990).
- [4] Jean-Pierre Aubin & Hélène Frankowska, *Set-Valued Analysis*, Birkhäuser Boston Inc., Cambridge, Mass., 1990.
- [5] Gerald Beer, "A geometric algorithm for approximating semicontinuous function," *Journal of Approximation Theory* 49 (1987), 31–40.

- [6] Gerald Beer, *Topologies on Closed and Closed Convex Sets*, Kluwer Academic Publishers, Dordrecht, 1993.
- [7] S. Boyd & L. Vandenberghe, “Semidefinite Programming,” *SIAM Review* 38 (1996), 49–95.
- [8] Charles Castaing & Michel Valadier, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Mathematics #580, Springer-Verlag, Berlin, 1977.
- [9] G. M. de Montricher, R. A. Tapia & J. R. Thompson, “Nonparametric maximum likelihood estimation of probability densities by penalty function methods,” *The Annals of Statistics* 3 (1975), 1329–1348.
- [10] Michael X. Dong, “Modeling prior information: a combined frequency and bayesian approach,” manuscript, University of California, Davis, 1999.
- [11] Michael X. Dong & Roger J-B Wets, “On the computation of the constrained maximum likelihood density estimator,” manuscript, University of California, Davis, 1999.
- [12] Michael X. Dong & Roger J-B Wets, “Convergence rates for the solutions of stochastic optimization problems,” manuscript, University of California, Davis, 1999.
- [13] Jitka Dupačová, “Epi-consistency in restricted regression models – The case of a general convex fitting function,” *Computational Statistics and Data Analysis* 14 (1992), 417–425.
- [14] Jitka Dupačová & Roger J-B Wets, “Asymptotic behavior of statistical estimators and of optimal solutions for stochastic optimization problems,” *The Annals of Statistics* 16 (1988), 1517–1549.
- [15] N. Etemadi, “An elementary proof of the strong law of large numbers,” *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 55 (1981), 119–122.
- [16] Stuart Geman & Chii-Ruey Hwang, “Nonparametric maximum likelihood estimation by the method of sieves,” *The Annals of Statistics* 10 (1982), 401–414.
- [17] Charles J. Geyer, “On the asymptotics of constrained M-estimation,” *The Annals of Statistics* 22 (1994), 1993–2010.
- [18] I. J. Good & R. A. Gaskins, “Nonparametric roughness penalties for probability densities,” *Biometrika* 58 (1971), 255–277.
- [19] P. Groeneboom, “Estimating a monotone density,” in *Proceedings of the Berkeley Conference in honor of J. Neyman and J. Kiefer* #2, Wadsworth Inc, 1985, 539–555.
- [20] P. Groeneboom & Jon A. Wellner, *Information Bounds and Nonparametric Maximum Likelihood Estimation*, Birkhäuser, Basel, 1992.

- [21] Christian Hess, “Lemme de Fatou et théorème de la convergence dominée pour des ensembles aléatoires non bornés et des intégrandes,” *Séminaire d’Analyse Convexe, Université du Languedoc* 16 (1986), 8.1–8.56.
- [22] Christian Hess, “Epi-convergence of sequences and integrands and strong consistency of the maximum likelihood estimator,” *Cahiers de Mathématiques de la Décision* no. 9121, Université Paris Dauphine, 1991.
- [23] R. Hettich & Ken O. Kortanek, “Semi-infinite programming: theory, methods and applications,” *SIAM Review* 35 (1993), 380–429.
- [24] Alan J. King & R. Tyrrell Rockafellar, “Asymptotic theory for solutions in statistical estimation and stochastic programming,” *Mathematics of Operations Research* 18 (1993), 148–162.
- [25] Alan J. King & Roger J-B Wets, “Epi-consistency of convex stochastic programs,” *Stochastics and Stochastics Reports* 34 (1990), 83–92.
- [26] V.K. Klonias, “Consistency of two nonparametric maximum penalized likelihood estimators of the probability density function,” *The Annals of Statistics* 10 (1982), 811–824.
- [27] Craig T. Lawrence, André L. Tits & Jian L. Zhou, “FSQP: A versatile tool for nonlinear programming user’s guide,” Manuscript, University of Maryland, College Park, 1994.
- [28] T. Leonard, “Density Estimation, stochastic processes and prior information,” *Journal of the Royal Statistical Society B* 40 (1978), 113–146.
- [29] Georg Pflug, “Asymptotic stochastic programs,” *Mathematics of Operations Research* 20 (1995), 769–789.
- [30] E. Polak & L. He, “A unified phase I - phase II method of feasible directions for semi-infinite optimization,” *Journal of Optimization Theory and Applications* 89 (1991), 83–107.
- [31] B.L.S. Prakasa Rao, *Nonparametric Functional Estimation*, Academic Press, 1983.
- [32] P. Révész, “Density estimation,” in *Handbook of Statistics*, P. Krishnaiah & P.K. Sen, eds. #4, North-Holland, Amsterdam, 1984, 531–549.
- [33] R.T. Rockafellar, “Integral functionals, normal integrands, and measurable selections,” in *Nonlinear Operators and the Calculus of Variations*, J.-P. Gossez & L. Waelbroeck, eds., Springer-Verlag Lecture Notes in Mathematics, no. 543, Berlin, 1976, 157–207.
- [34] R.T. Rockafellar & Roger J-B Wets, “Variational systems, an introduction,” in *Multifunctions and Integrands: Stochastic Analysis, Approximation and Optimization*, G. Salinetti, ed., Springer Verlag Lecture Notes in Mathematics 1091, Berlin, 1984, 1–54.
- [35] Werner Römisch & Rüdiger Schultz, “Stability analysis for stochastic programs,” *Annals of Operations Research* 30 (1991), 241–266.

- [36] Gabriella Salinetti, “Alla base dei rapporti fra Probabilità e Statistica: processi empirici e funzionali statistici,” in *Società Italiana di Statistica atti della XXXV Riunione Scientifica* #1, CEDAM, Padova, 1990, 75–96.
- [37] Gabriella Salinetti & Roger J-B Wets, “On the convergence in distribution of measurable multifunctions (random sets), normal integrands, stochastic processes and stochastic infima,” *Mathematics of Operations Research* 11 (1986), 385–419.
- [38] Francisco J. Samaniego & Dana M. Reneau, “Toward a reconciliation of the bayesian and frequentist approaches to point estimation,” *Journal of the American Statistical Association* 89 (1994), 947–957.
- [39] Eugene F. Schuster, “Incorporating support constraints into nonparametric estimators of densities,” *Communications in Statistics-Theory and Methods* 14 (1985), 1123–1136.
- [40] Alexander Shapiro, “Asymptotic behavior of optimal solutions in stochastic programming,” *Mathematics of Operations Research* 18 (1993), 829–845.
- [41] Frank H. Shaw & Charles J. Geyer, “Estimating and testing in constrained covariance models,” manuscript, University of Minnesota, 1995.
- [42] Xiaotong Shen & Wing Hung Wong, “Convergence rate of sieve estimates,” *The Annals of Statistics* 22 (1994), 580–615.
- [43] B. W. Silverman, “On the estimation of a probability density function by the maximum penalized likelihood method,” *The Annals of Statistics* 10 (1982), 795–810.
- [44] James R. Thompson & Richard A. Tapia, *Nonparametric Function Estimation, Modeling, and Simulation*, SIAM, Philadelphia, 1990.
- [45] Sara Van de Geer, “A new approach to least squares estimation,” Manuscript, Center for Mathematics and Computer Science, Amsterdam, 1987.
- [46] Wim Vervaat, “Random upper semicontinuous functions and extremal processes,” Report MS-R8801, Center for Wiskunde en Informatica, Amsterdam, 1988.
- [47] Abraham Wald, “Note on the consistency of the maximum likelihood estimate,” *Annals of Mathematical Statistics* 20 (1949), 595–601.
- [48] Jane-Ling Wang, “Strong consistency of approximate maximum likelihood estimators with applications in nonparametrics,” *The Annals of Statistics* 13 (1985), 932–946.
- [49] Jinde Wang, “Limit distribution of L^1 -estimators for constrained nonlinear regression problems,” manuscript, University of Nanjing.
- [50] Roger J-B Wets, “Laws of large numbers for random lsc functions,” *Applied Stochastic Analysis & Stochastics Monographs* 5 (1991), 101–120.
- [51] Roger J-B Wets, “Constrained estimation: consistency and asymptotics,” *Applied Stochastic Models and Data Analysis* 7 (1991), 17–32.

- [52] R.A. Wijsman, “Convergence of sequences of convex sets, cones and functions. II,” *Transactions of the American Mathematical Society* 123 (1966), 32–45.